Biologia Molecular Computacional
IBI5035/QBQ2507 - 2023

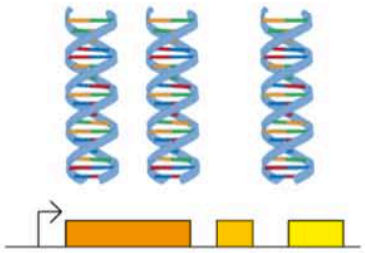
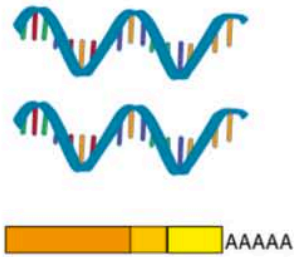
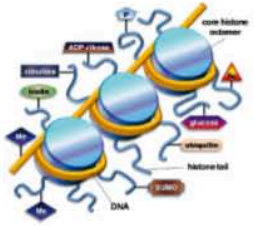
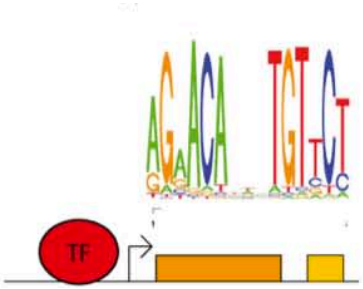
Bancos de Dados Biológicos

Eduardo Moraes Rego Reis
Instituto de Química - USP

Estudos ômicos geram grande quantidade de dados moleculares

DNA → RNA → Proteína

modificação química no DNA/RNA/histonas

Genômica	Transcritômica	Proteômica	Epigenômica
		 > Replicase polyprotein 1ab, P19811 MATFSATGFG GSFVRDWSLD LPDACE- HGAG LCCEVDGSTL CAECFRGCEG M	 
· Sequenciamento de DNA	- Microarranjos - RNA-Seq	· Espectrometria de massas	· ChIP-Seq · Sequenciamento DNA bissulfito	· ChIP–Seq
Sequencia de genes e genomas	· Sequencia de RNAs · Abundância de RNAs	· Sequencia, abundância, modificações pós-tradução de proteínas	· Modificações de histonas, metilação de DNA	· Sítios de ligação de fatores de transcrição · Estados de ativação da cromatina

Tecnologias de células únicas

http://en.wikipedia.org/wiki/List_of_biological_databases

Contents [hide]

Categorias de bancos de dados

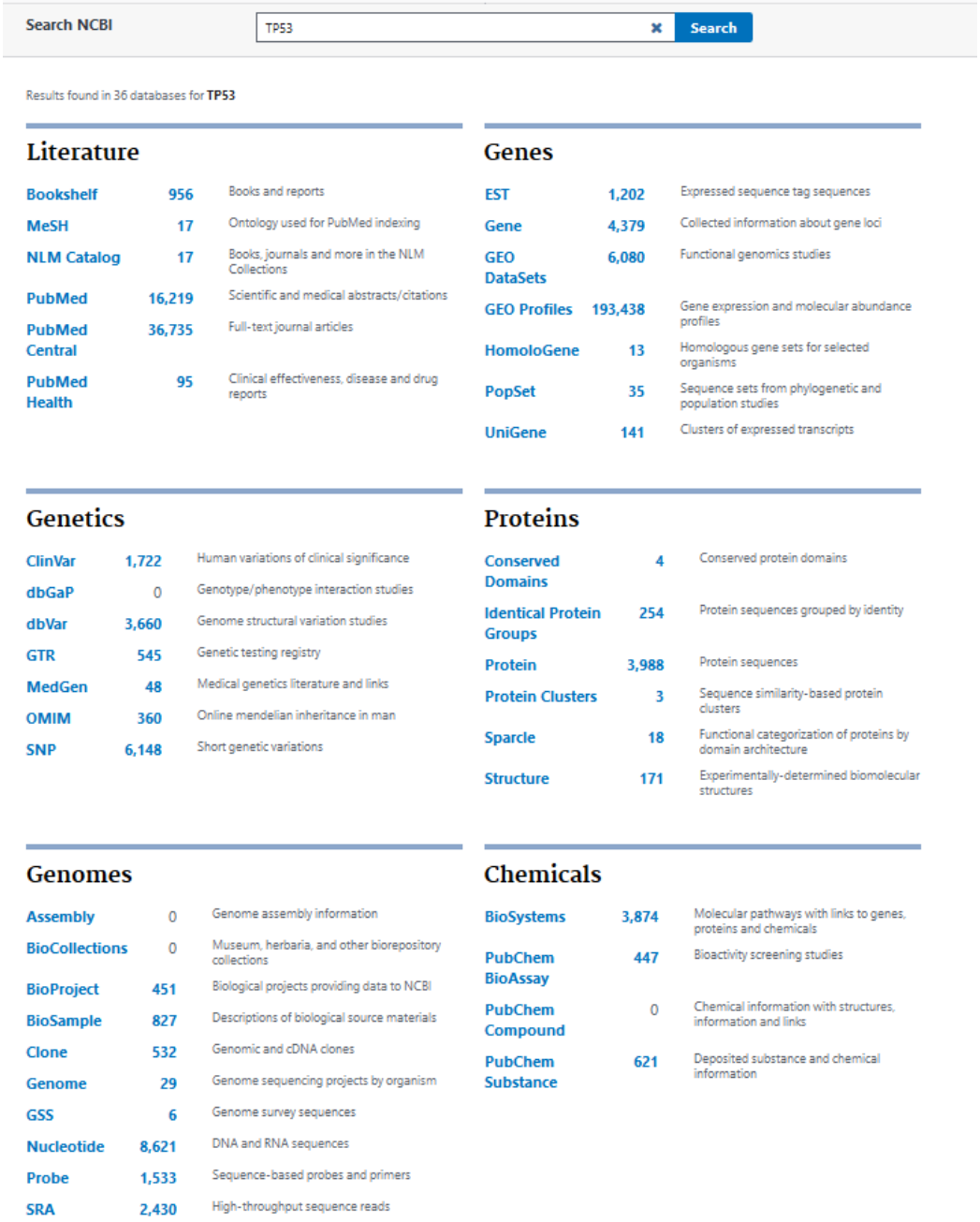
Primários
Dados gerados experimentalmente
ex. sequencias/estruturas de DNA, RNA, proteínas.
Dados genomicos brutos e normalizados de RNAseq “bulk”e celula única

Função de arquivamento.

Secundários
dados resultantes da análise, interpretação e curagem de dados primários
ex. famílias de proteínas, genes alvos de microRNAs, variação/regulação/função de sequencias de DNA (SNPs, SNVs, Indels)

Agregação de informação

http://www.ncbi.nlm.nih.gov/Entrez/



ngdc.cncb.ac.cn/cancerscem/index

ChatGPT: Optimiz... | Portal de Serviços... | Projeto pagina da... | Tematico melanoma | Grupos na Italia | Pedro Sebes | Aulas | Tematico FAPESP... | Todos os favorito

CNCB NGDC

Databases Tools Standards Publications Abo

Cancer Single-cell Expression Map

A public database dedicating to collecting, analyzing, visualizing single-cell RNA-Seq data of human cancers. Multi-level analyses were performed to deeply explore the tumor microenvironment of different types of human cancers and a comprehensive online analyze platform was equipped in the database.

Home | Project Browse | Search | Analyze | Documentation | References | Downloads | Help

Data release 1.0

Quick Search Please enter 1 or more characters

(e.g.Lung Adenocarcinoma, *TP53*, ENSG00000186891)

28 PROJECTS

20 CANCER TYPES

638,341 CELLS

Keyword Cloud

GEPIA2

Gene Expression Profiling Interactive Analysis

Single Gene Analysis | Cancer Type Analysis | Custom Data Analysis | Multiple Gene Analysis

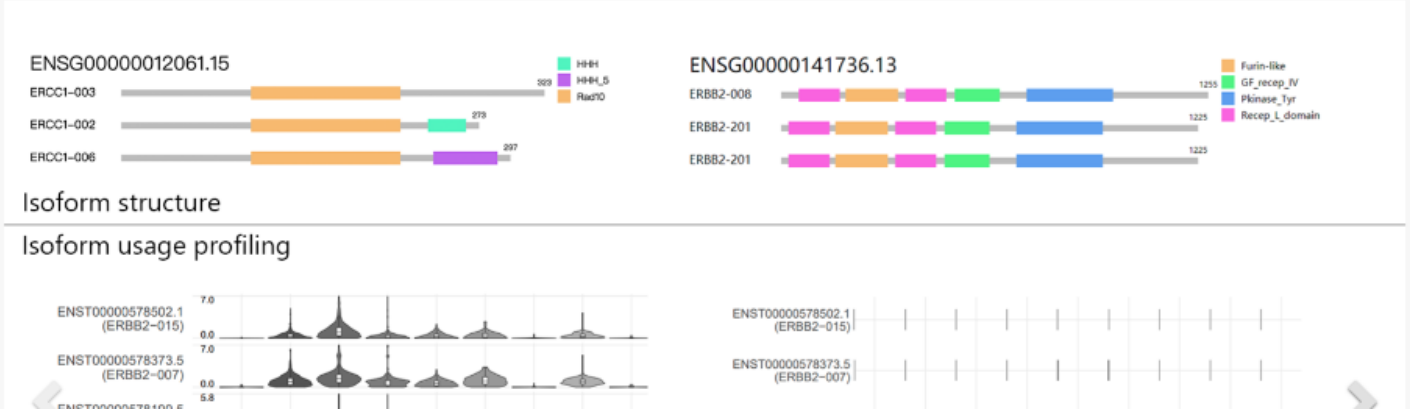
Enter gene/isoform name:

The indicators in search box are "symbol" or "alias (newest symbol)".

e.g. ERBB2/ENSG00000141736 or ERBB2-001/ENST00000584601

GoPIA!

Profile | Boxplots | Stage Plots | Survival Analysis | Similar



Bancos de dados para anotação, visualização e integração de dados genômicos com informações da literatura

Ex. UCSC Genome Browser, ENSEMBL

Tipos de dados representados em visualizadores de dados genômicos

Estrutura e padrão de expressão gênica

- isoformas de splicing
- estrutura exon/intron
- regiões codificadoras/UTRs
- expressão em diferentes tecidos/células

Regiões regulatórias no DNA

- Promotores, enhancers
- sítios de ligação de fatores de transcrição, histonas

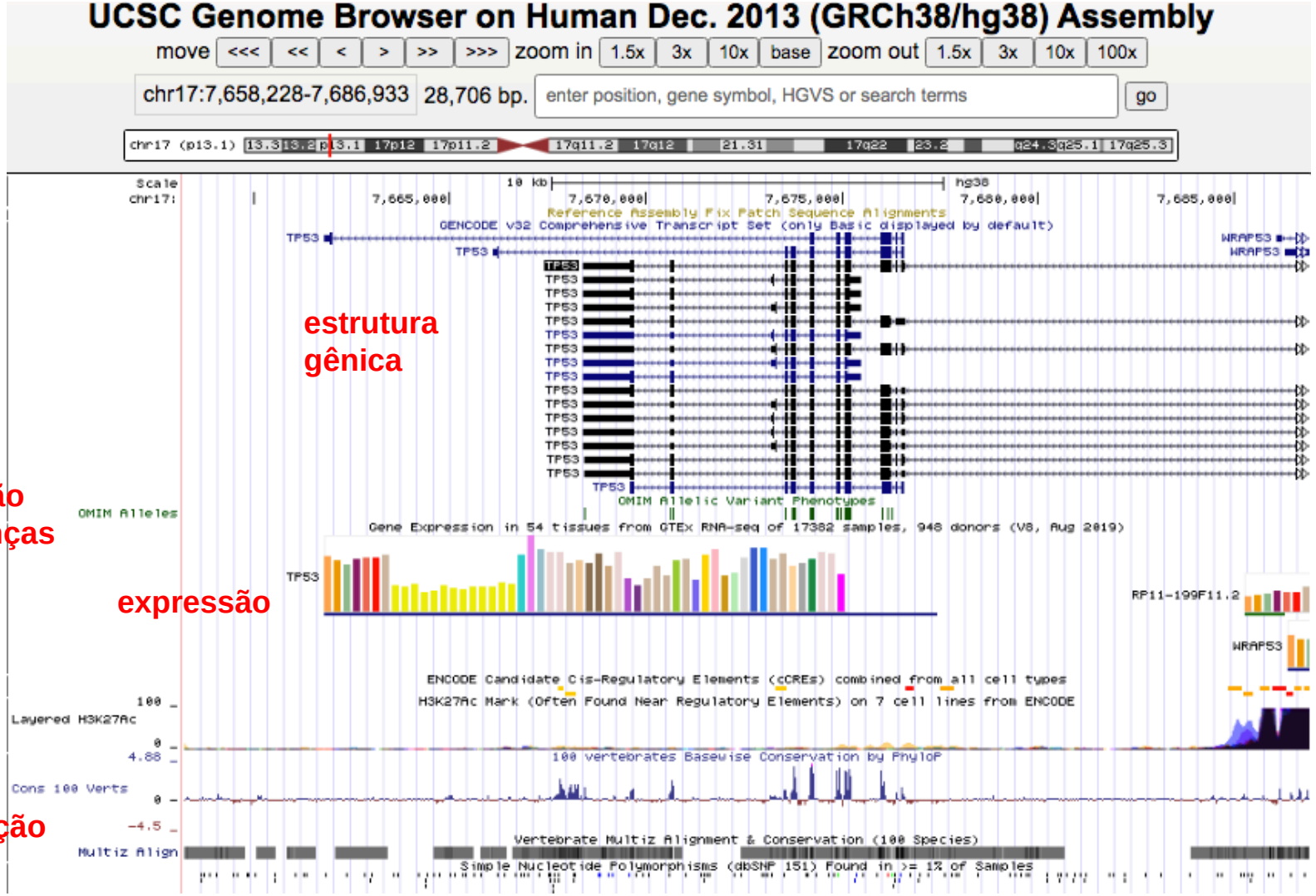
Variação na sequencia do DNA/RNA

- SNPs, indels,
- mutações, edição de RNA

Conservação evolutiva

alinhamento de DNA entre espécies

http://genome.ucsc.edu/cgi-bin/hgGateway

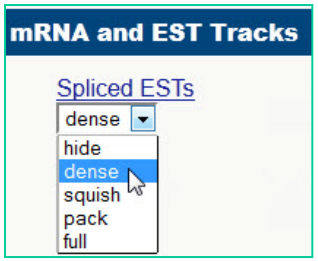


Annotation Tracks

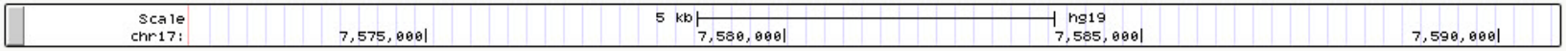
Informação está organizada em tracks de anotação

- chromosome band
- gap locations
- known genes
- predicted genes
- phenotype and disease
- enhancer/promoter data
- microarray/expression data
- evolutionary conservation
- SNPs and structural variation
- repeated regions
- more…

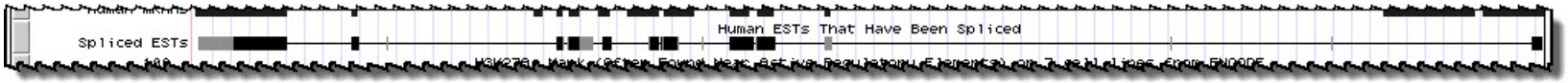
Como controlar a visibilidade das tracks



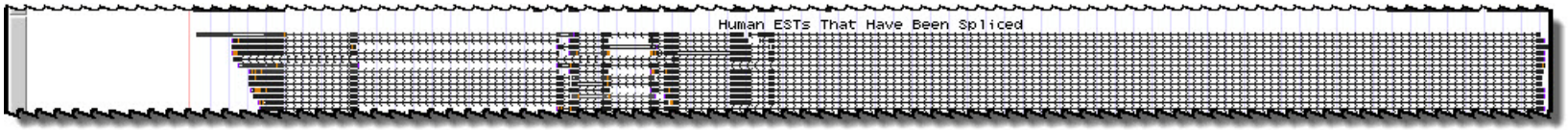
- Hide: removes a track from view



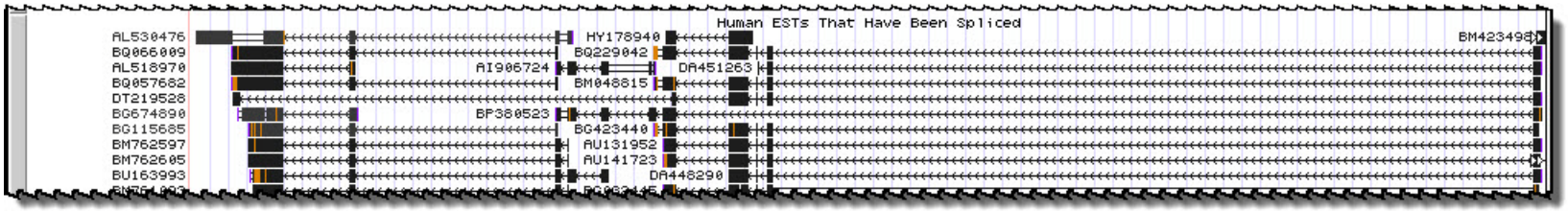
- Dense: all items collapsed into a single line



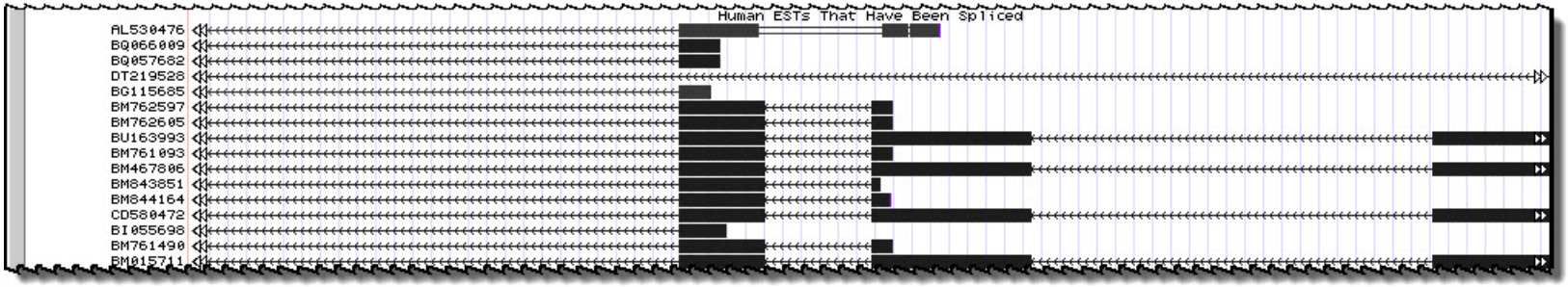
- Squish: each item = separate line, but 50% height + packed



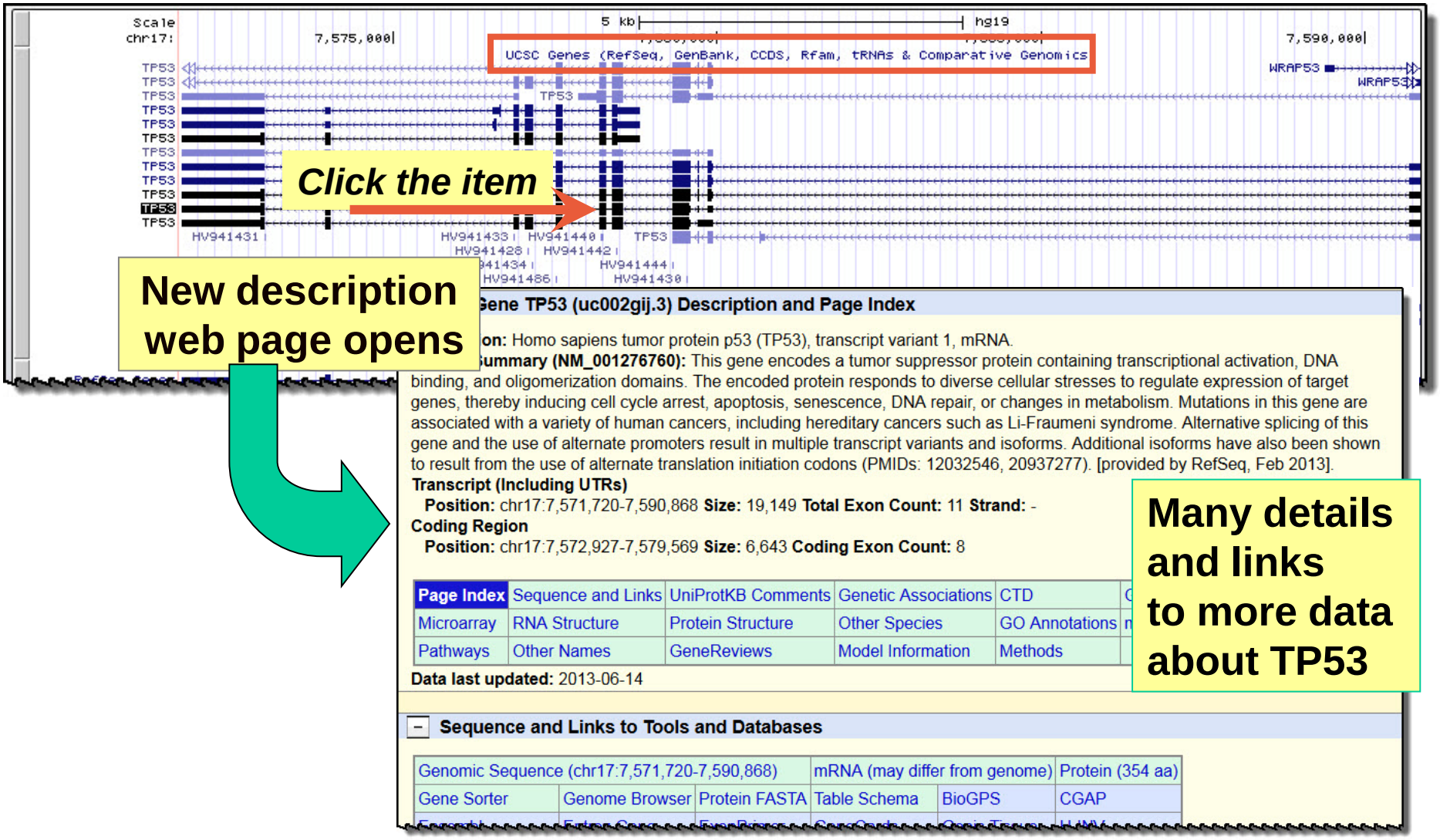
- Pack: each item separate, but efficiently stacked (full height)



- Full: each item on separate line (may need to zoom to fit)



Os tracks estão conectados com diferentes camadas de informação



Example: click your mouse anywhere on the TP53 line

Formatos para representação de dados genômicos

http://genome.ucsc.edu/FAQ/FAQformat.html#format1

General formats

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigNarrowPeak table format
- bigWig format
- Chain format
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- interact and bigInteract format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

- **Desafio: Dados de diferentes naturezas: estrutura gênica, regulação, expressão, variação, etc.**

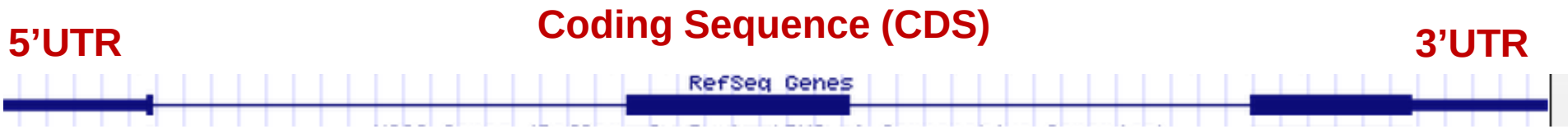
Em comum:
- **contem linhas de colunas separadas por tabulação**
- **baseados em coordenadas genômica**

Formato BED (Browser Extensible Data)

BED 3, **BED 6**, BED 12

chr	chromStart	chromEnd	name	score	strand
chr1	67051159	67163158	NM_024763	0	-
chr1	67075869	67163158	NM_207014	0	-
chr1	50286272	50440127	NM_001144777	0	+
chr1	41744622	42156965	NM_024503	0	-
chr1	8335050	8800286	NM_001042681	0	-
chr1	33544953	33559286	NM_001080438	0	-
chr1	8335050	8406334	NM_001042682	0	-
chr1	8335050	8800286	NM_012102	0	-
chr1	41744622	42156965	NM_001127714	0	-
chr1	58718978	58785034	NM_145243	0	-
chr1	16761508	16812687	NM_017940	0	-
chr1	100590610	100723514	NM_033313	0	+
chr1	100590610	100737609	NM_033312	0	+
chr1	100590610	100758421	NM_003672	0	+
chr1	75444662	75849387	NM_152697	0	-
chr1	117404471	117447014	NM_003594	0	+
chr1	75440403	75849387	NM_001130058	0	-
chr1	92268120	92301681	NM_173567	0	+
chr1	184532027	184550311	NM_001127709	0	+
chr1	184532027	184550311	NM_001127710	0	+
chr1	184532027	184550311	NM_005807	0	+
chr1	184532027	184550311	NM_001127708	0	+
chr1	184547408	184611080	NM_003292	0	-
chr1	167747815	167822393	NM_000130	0	-
chr1	243199793	243317771	NM_032328	0	+
chr1	243200253	243317771	NR_026588	0	+
chr1	243199906	243355153	NR_026586	0	+
chr1	243200253	243355153	NR_026587	0	+
chr1	226462453	226615574	NM_052843	0	+
chr1	226462453	226633198	NM_001098623	0	+
chr1	243200253	243355153	NM_001143943	0	+
chr1	226462453	226633198	NM_001271223	0	+
chr1	6246918	6341591	NM_181866	0	-
chr1	6767970	7752353	NM_015215	0	+
chr1	2026014	2106694	NM_001033582	0	+

Arquivo BED com coordenadas do gene MYC




```
chr8    128817496    128822862    NM_002467    0    +    128818021    128822386    0    3    555,772,1039,    0,2179,4327,
```

campo 1: chr8	cromossomo
campo 2: 128817496	coordenada início do gene
campo 3: 128822862	coordenada fim do gene
campo 4: NM_002467	nome do gene (GeneID)
campo 5: 0	score (definição da representação gráfica)
campo 6: +	fita em que o gene se localiza
campo 7: 128818021	coordenada início da CDS
campo 8: 128822386	coordenada fim da CDS
campo 9: 0	itemRGB (definição da cor no gráfico)
campo 10: 3	número de exons
campo 11: 555,772,1039,	tamanho dos exons (em nt)
campo 12: 0,2179,4327,	coordenada de início dos exons (em relação ao campo 2)

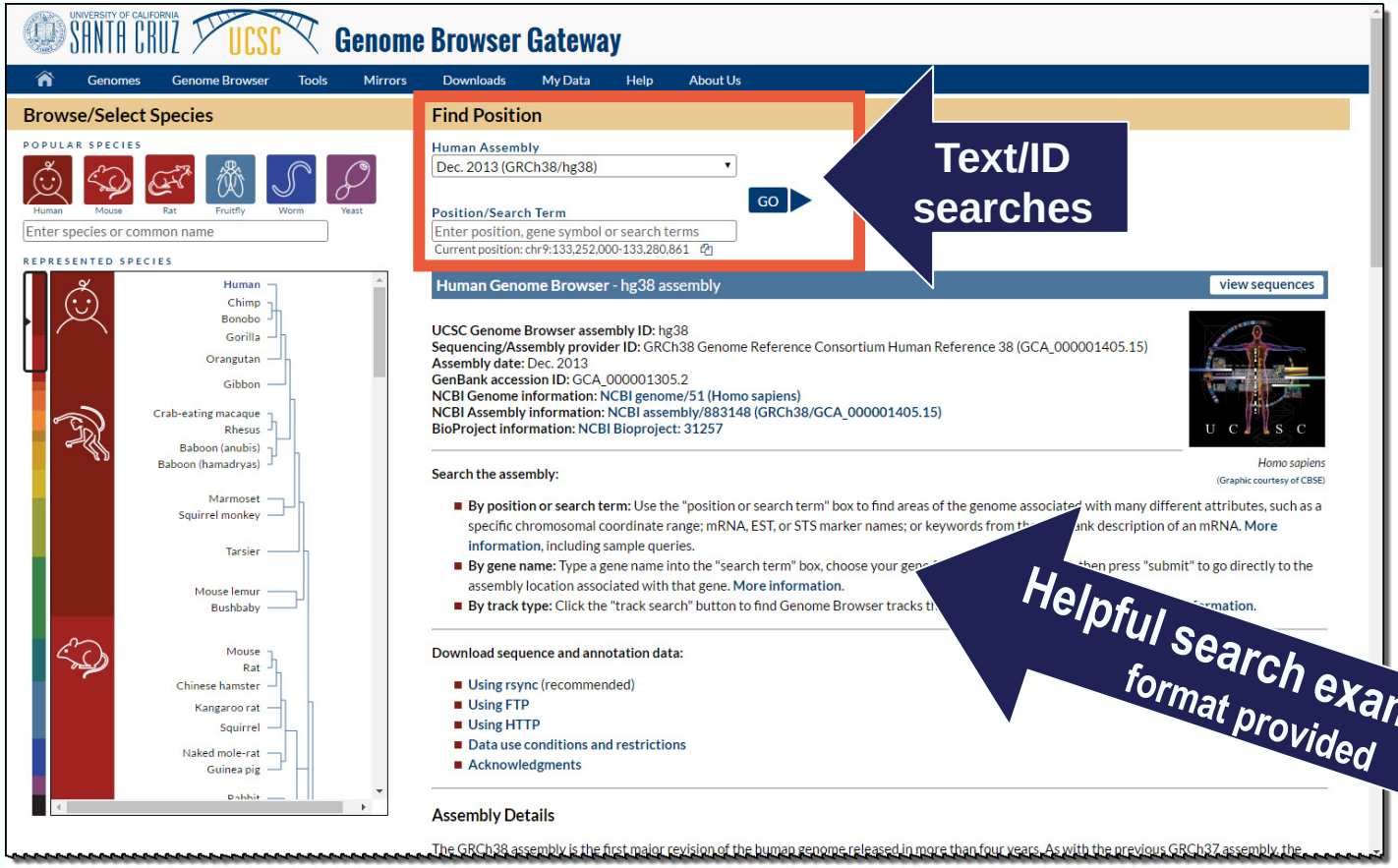
http://genome.ucsc.edu/cgi-bin/hgGateway

Através do Genome Browser é possível:

- Fazer consultas textuais utilizando o nome, símbolo, localização genômica
- Alinhar sequencias de nucleotídeo/proteína no genoma (BLAT)
- Cruzar informações de “tracks” de anotação e selecionar regiões de DNA de interesse (“Table Browser")

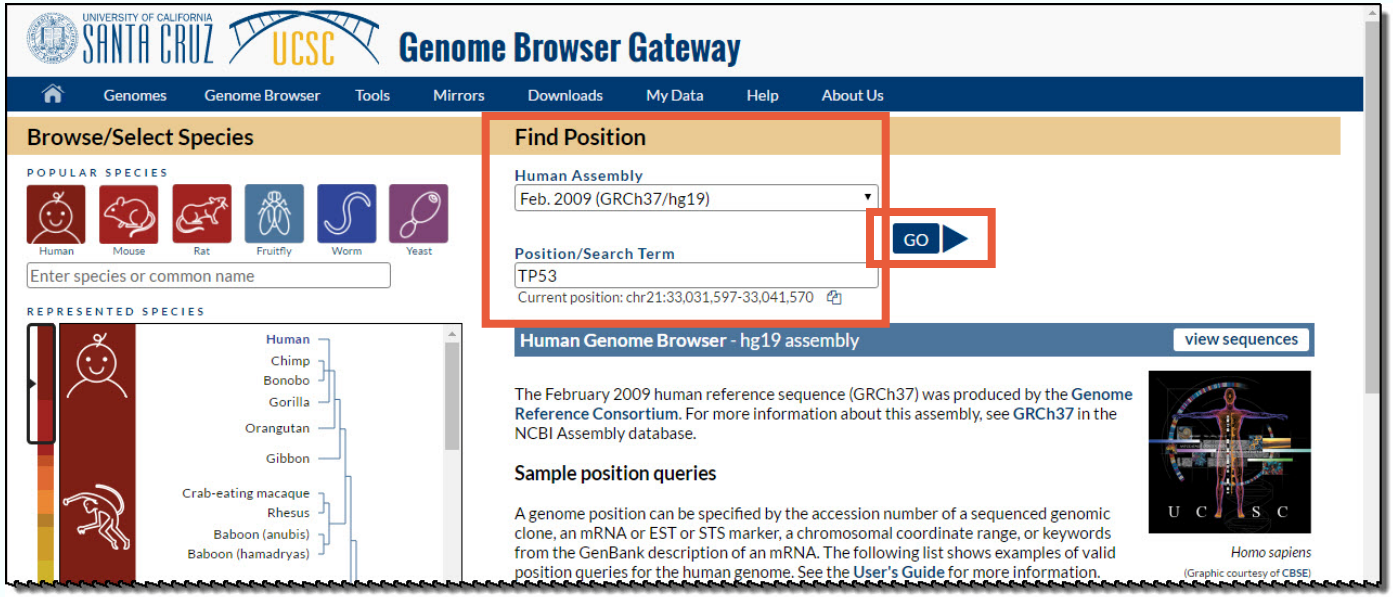
consultas textuais utilizando o nome, símbolo, localização genômica

Gateway: Start Page for a Basic Search

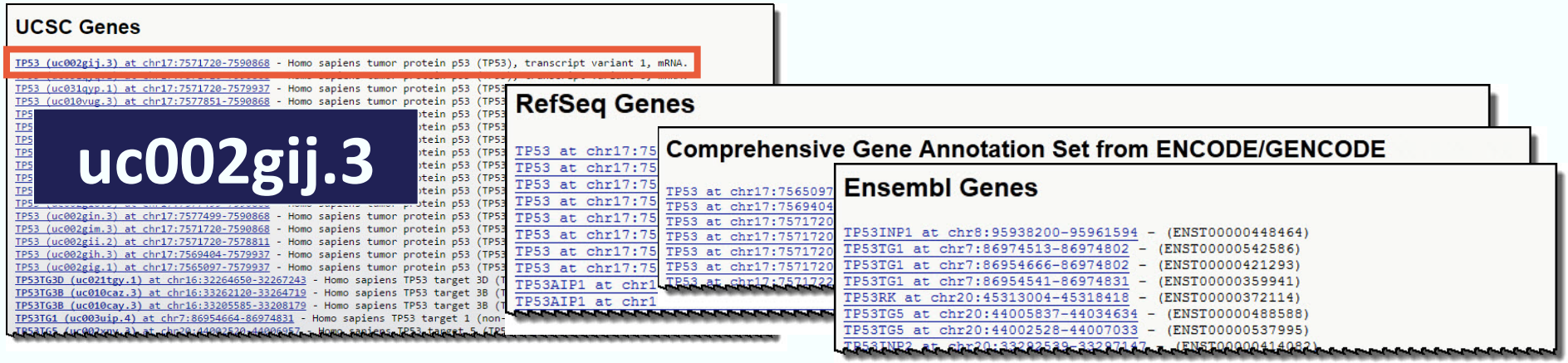


- Use this Gateway to search:
  - Gene names, symbols, IDs
  - Chromosome number: chr7, or region: chr11:1038475-1075482
  - Keywords: kinase, receptor

Sample Search for Human TP53

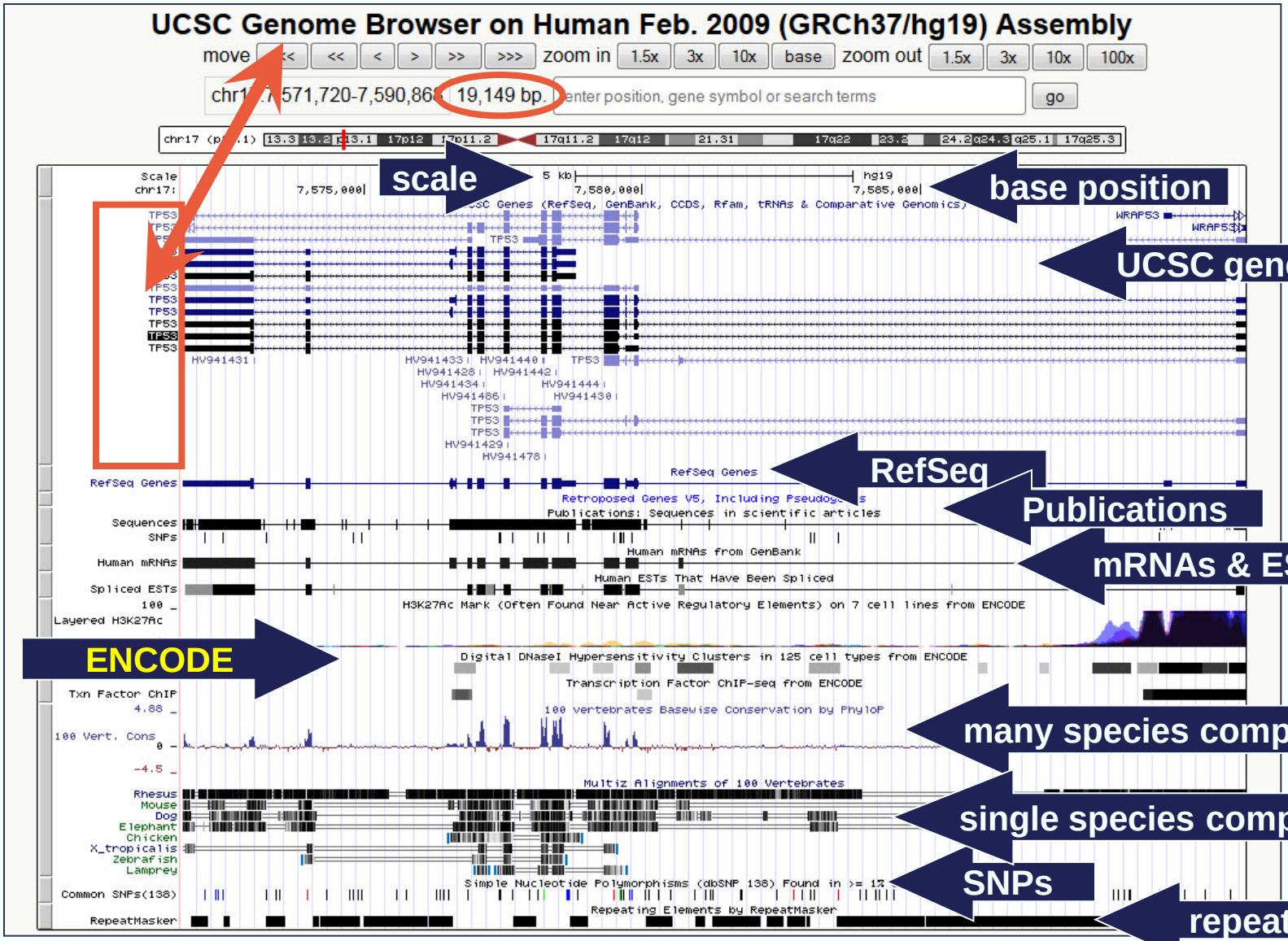


- Sample search: human, February 2009 assembly, tp53



- Select from results list; or goes to a viewer page, if unique

Sample Genome Viewer Image, TP53 Region



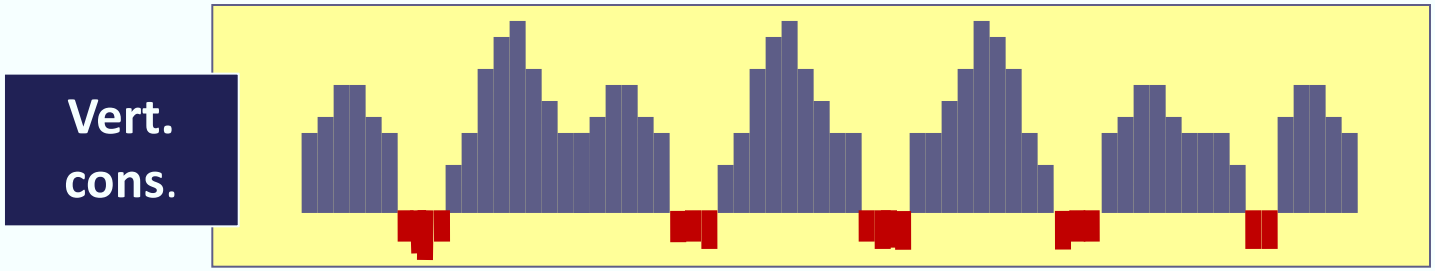
Visual Cues on the Genome Browser

Tick marks; a single location (STS, SNP)

Intron and direction of transcription <<< or >>>

Track colors *may* have meaning—for example, UCSC Gene track:

- If there is a corresponding PDB entry = black
- If there is a corresponding reviewed/validated seq = dark blue
- If there is a non-RefSeq seq = lightest blue

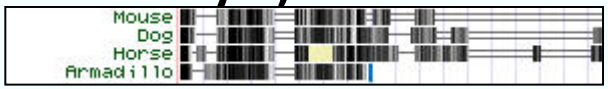


Wiggle

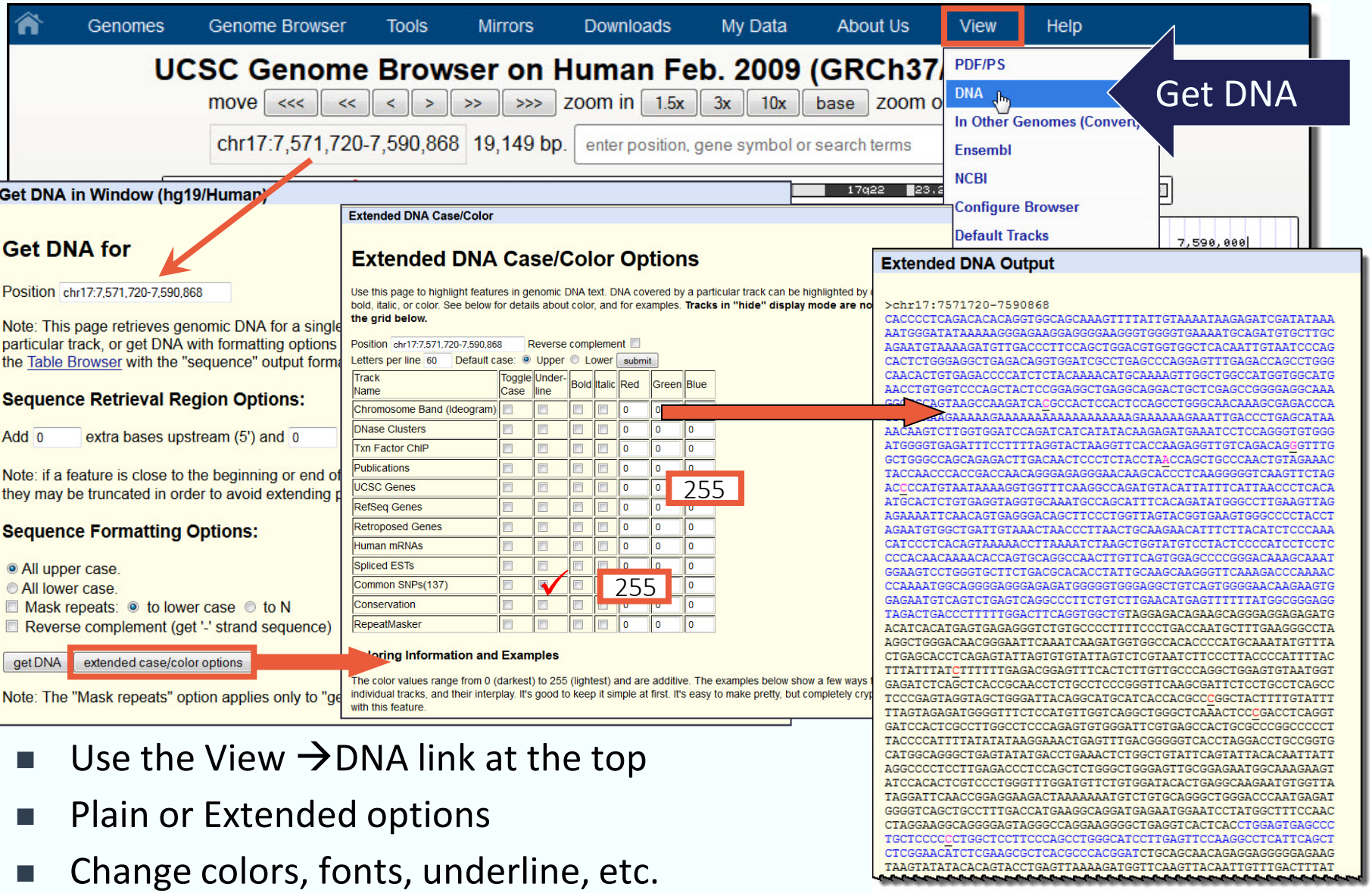
height of a blue bar is increased likelihood of conservation,
red indicates a likelihood of faster-evolving regions

Alignment indications (Conservation pairs: "chain" or "net" style)

- Alignments = boxes, Gaps = lines



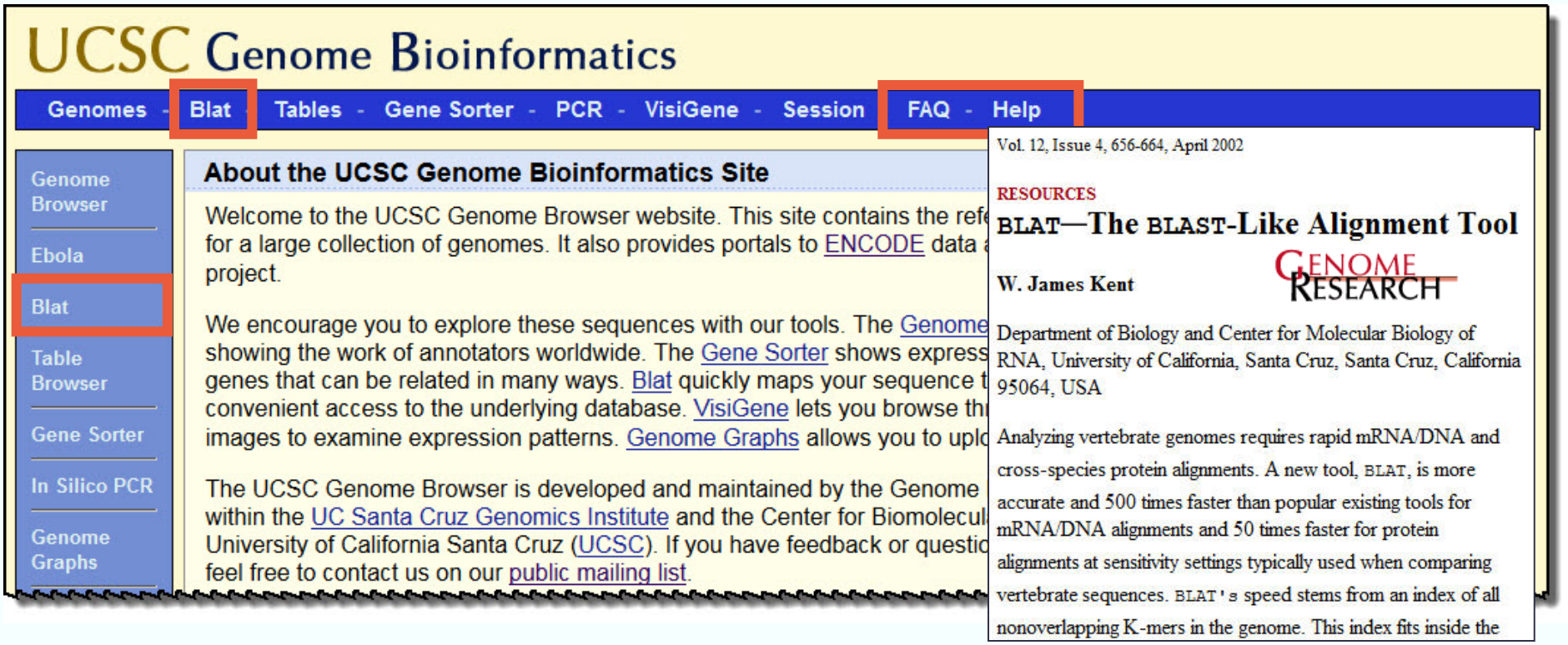
Get DNA, with Extended Case/Color Options



- Use the View →DNA link at the top
- Plain or Extended options
- Change colors, fonts, underline, etc.

Alinhar sequencias de nucleotídeo/proteína no genoma (BLAT)

Accessing the BLAT Tool

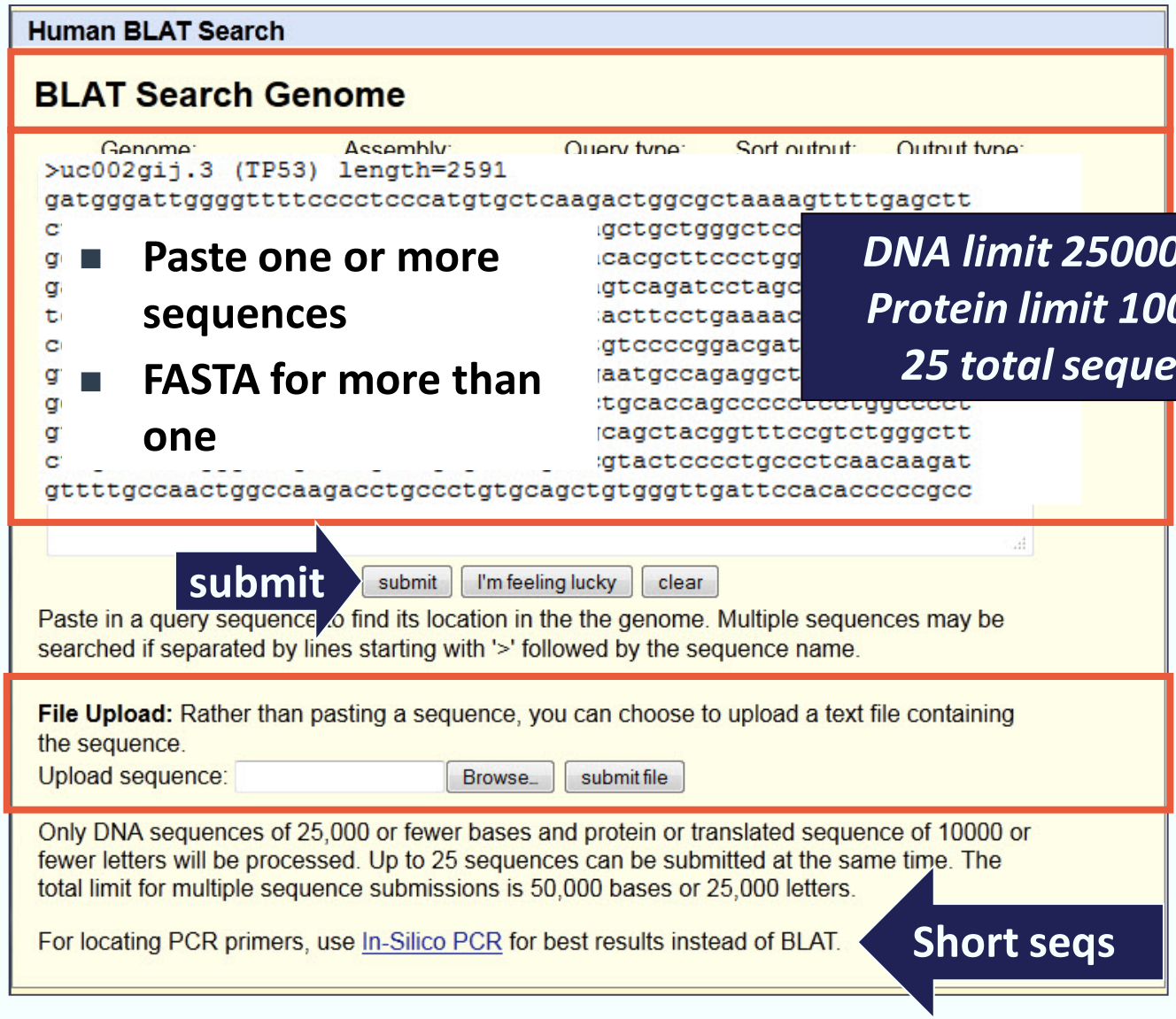


BLAT = BLAST-like Alignment Tool

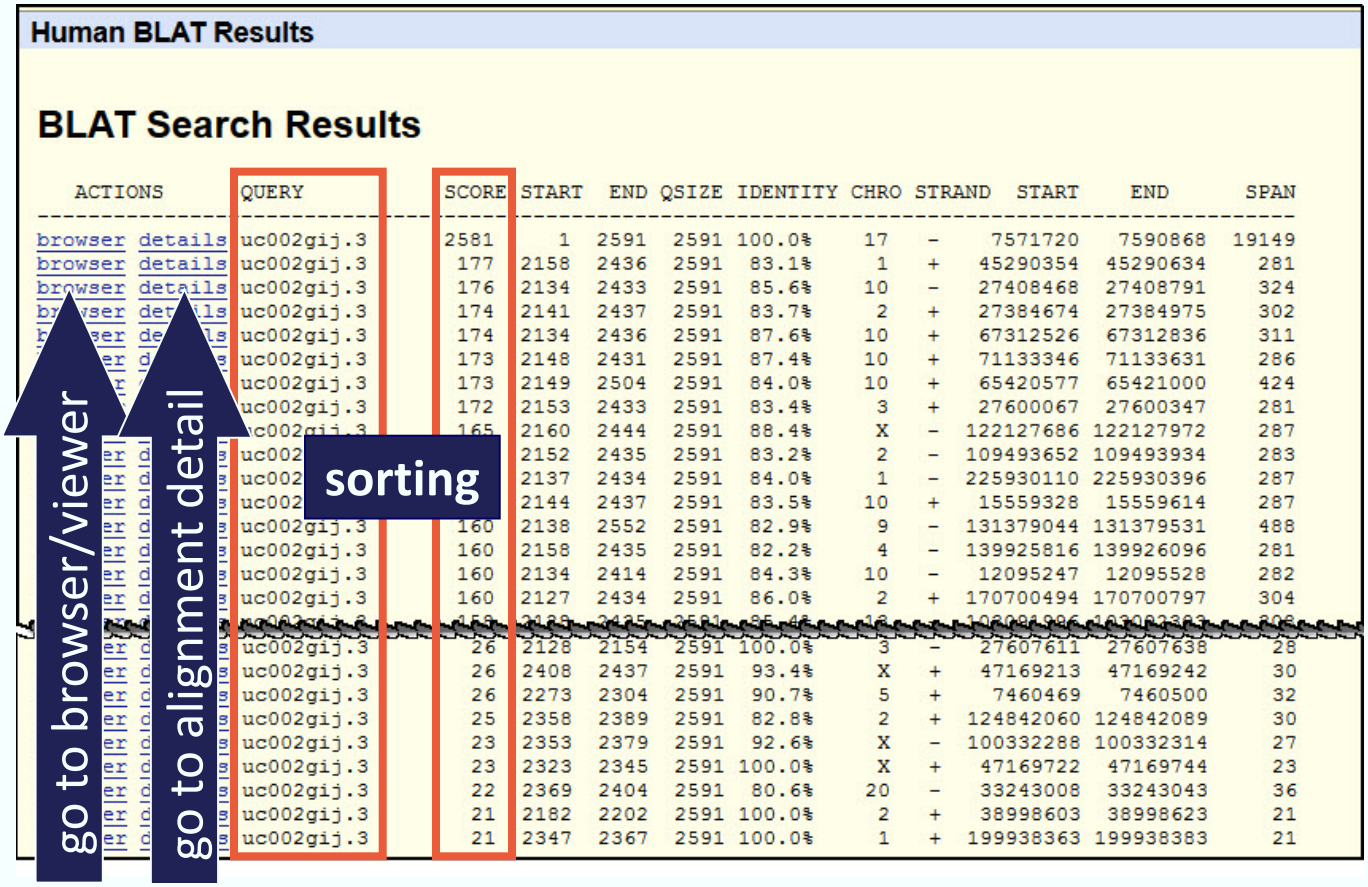
- Rapid searches by INDEXING the entire genome
- Works best with high similarity matches
- See documentation and publication for details
  - Kent, WJ. *Genome Res*. 2002. 12:656 and “Help”

BLAT Tool Interface

- Make choices
- Or upload

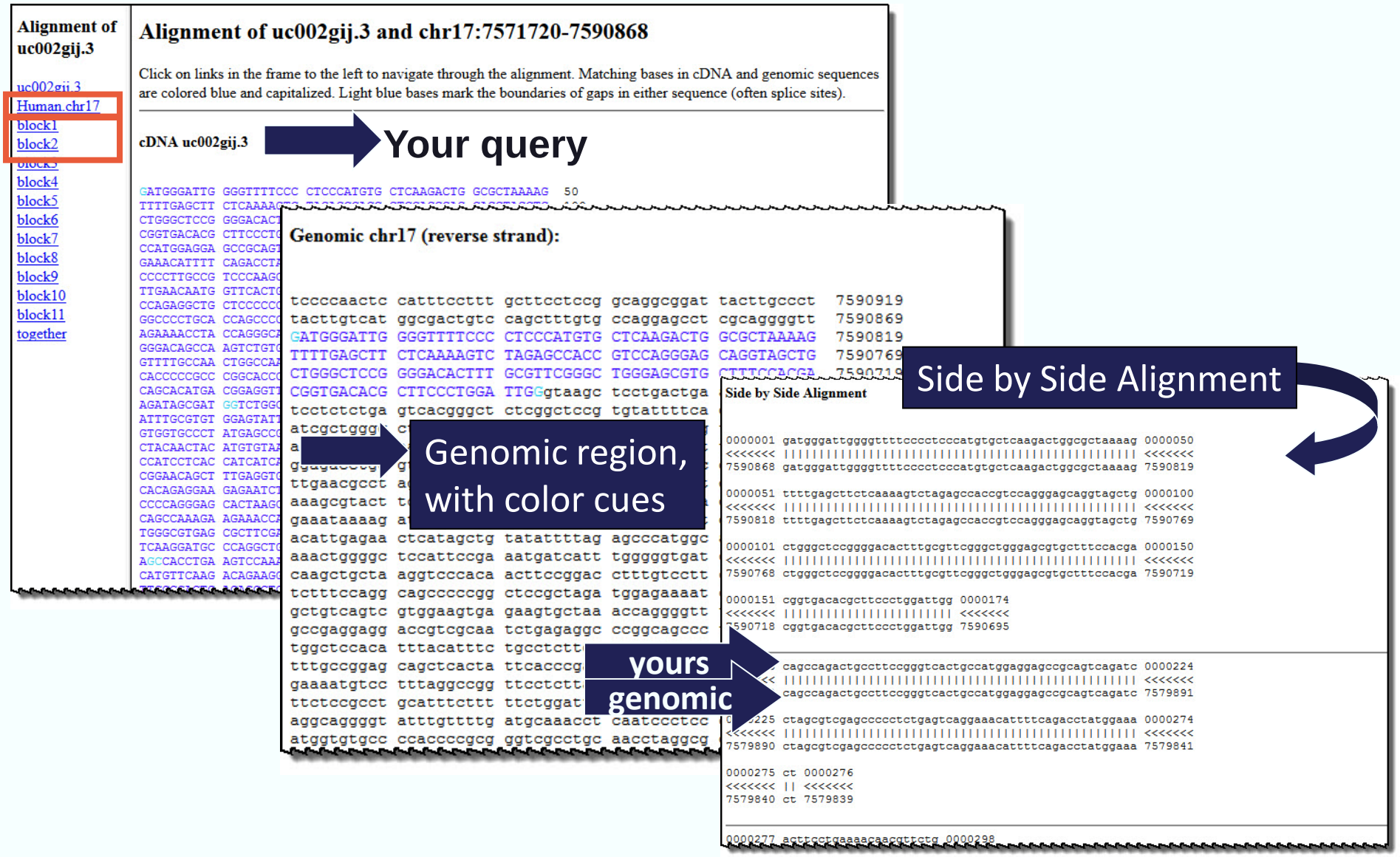


BLAT Results with Hyperlinks



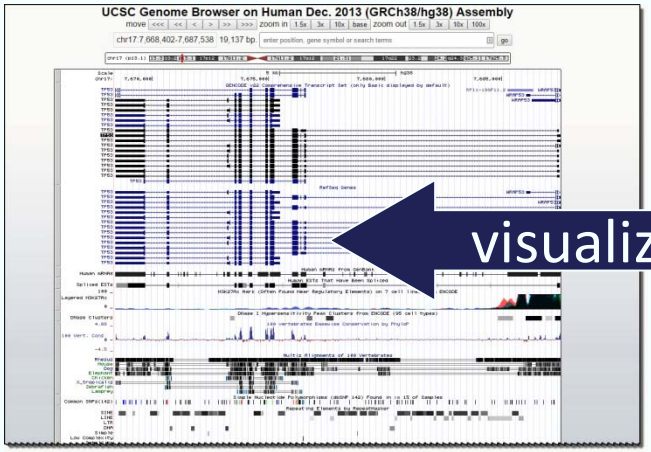
- Results with demo sequences, settings default; sort = Query, Score
  - **Score is a count of matches—higher number, better match**
- Click browser to go to Genome Browser image location (next slide)
- Click details to see the alignment to genomic sequence (2nd slide)

BLAT Results, Alignment Details



Cruzar informações de tracks de anotação e selecionar regiões de DNA de interesse (Table Browser)

Genome Browser Database



visualize

Underlying Database (MySQL)

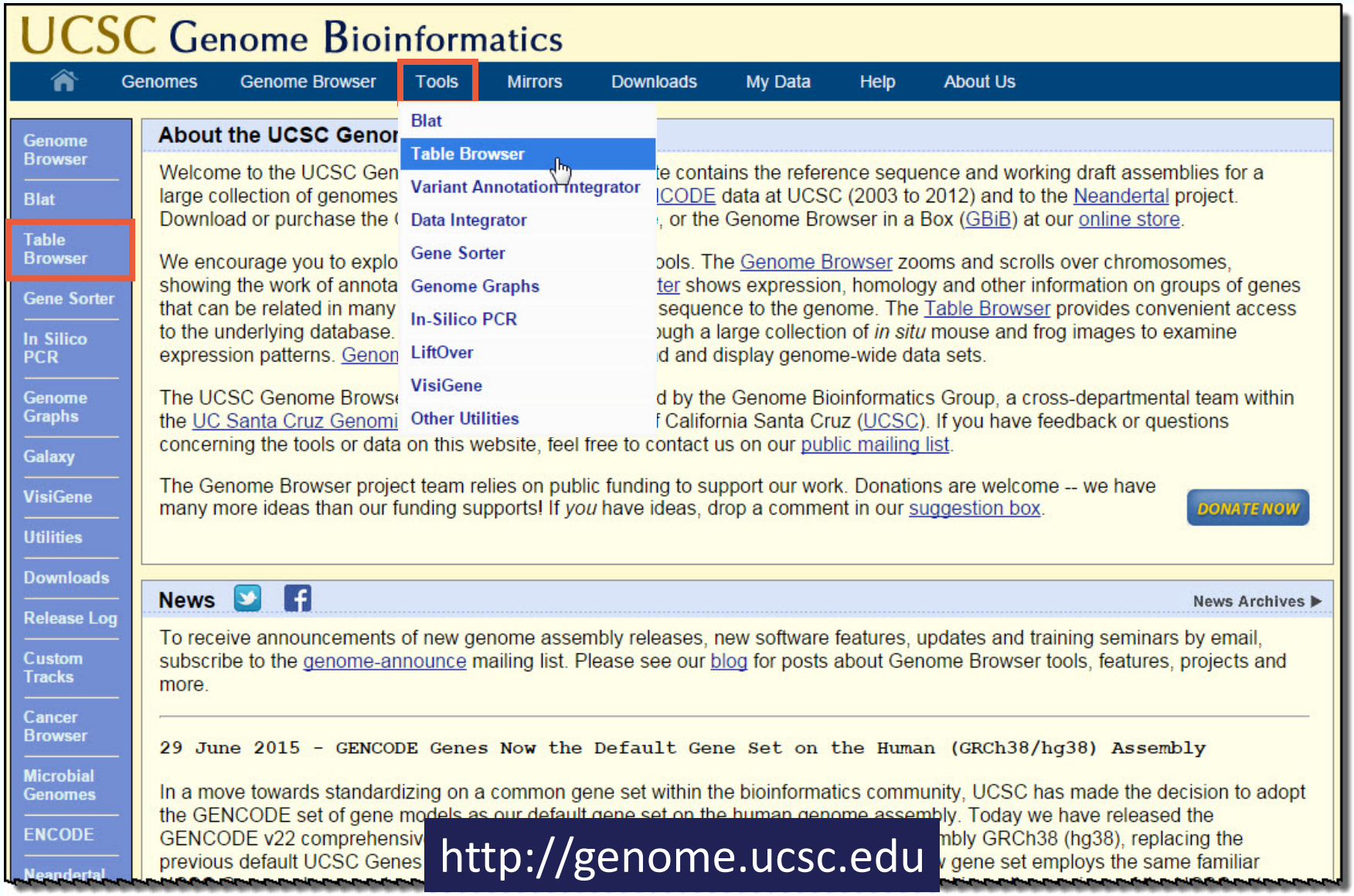
search & download

#chrom	chromStart	chromEnd	name	class	func
chr17	38451241	38451242	rs28897699	single	coding-nonsynon,untranslated,cd
chr17	38451354	38451355	rs8176316	single	untranslated,intron
chr17	38452072	38452073	rs8176315	single	untranslated,intron
chr17	38452146	38452147	rs8176314	single	untranslated,intron
chr17	38452200	38452201	rs9896925	single	untranslated,intron
chr17	38452527	38452528	rs8176313	single	untranslated,intron
chr17	38452648	38452649	rs8176312	single	untranslated,intron
chr17	38452664	38452665	rs8176311	single	untranslated,intron
chr17	38452703	38452704	rs8068463	single	untranslated,intron
chr17	38453219	38453220	rs4438367	single	coding-synon,untranslated,cds-r
chr17	38453227	38453228	rs28897698	single	coding-nonsynon,untranslated,cd
chr17	38453438	38453439	rs8176310	single	untranslated,intron
chr17	38453634	38453635	rs4793190	single	untranslated,intron
chr17	38453634	38453635	rs7406700	single	untranslated,intron
chr17	38454595	38454596	rs8176307	single	untranslated,intron
chr17	38454870	38454871	rs8176306	single	untranslated,intron
chr17	38454889	38454890	rs8176305	single	untranslated,intron
chr17	38454957	38454958	rs8176304	single	untranslated,intron
chr17	38454990	38454991	rs8176303	single	untranslated,intron
chr17	38455227	38455228	rs3092988	single	untranslated,intron
chr17	38455362	38455363	rs8176302	single	untranslated,intron
chr17	38455548	38455549	rs8176301	single	untranslated,intron
chr17	38455820	38455821	rs8176300	single	untranslated,intron
chr17	38455966	38455967	rs8176299	single	untranslated,intron
chr17	38456104	38456105	rs12949768	single	untranslated,intron
chr17	38456115	38456116	rs11079053	single	untranslated,intron

Primary table: positions, names, etc.

Auxiliary table: related data

The Table Browser



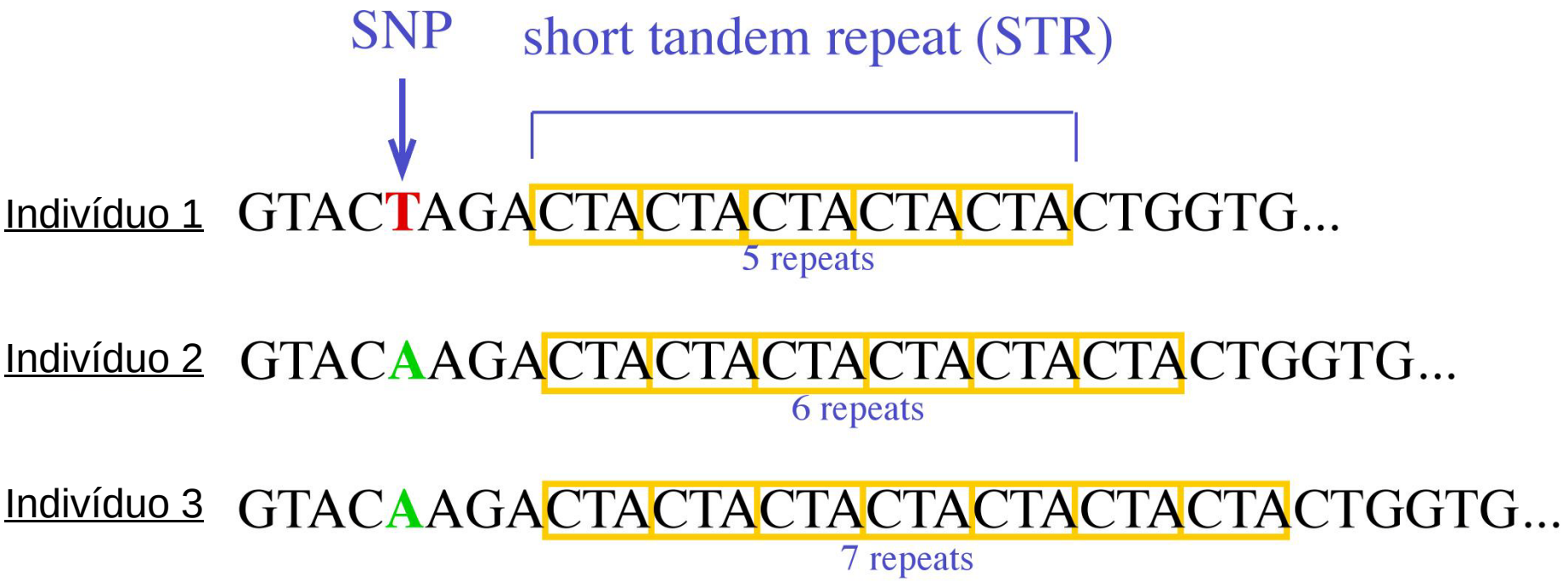
Tipos de polimorfismos frequentemente encontrados nos genomas

- Variações em uma única base:

Single Nucleotide Polymorphisms, **SNPs**

- Variações que afetam múltiplas bases no genoma:

Ex. Diferenças no número de trechos repetitivos no DNA



Discovery using the Table Browser

OpenHelix
A Division of BioITWorld

Table 1 | **Features of trinucleotide expansion in humans**

Disease	Sequence	Location	Parent of origin of expansion	Repeat number (normal)	Repeat number (pre-mutation)	Repeat number (disease)	Somatic instability
Diseases with coding TNRs							
DRPLA	CAG	*ATN1* (exon 5)	P	6–35	35–48	49–88	Yes
HD	CAG	*HTT* (exon 1)	P	6–29	29–37	38–180	Yes
OPMD	GCN	*PABPN1* (exon 1)	P and M	10	12–17	>11	None found in tissue tested (hypothalamus)
SCA1	CAG	*ATXN1* (exon 8)	P	6–39	40	41–83	Yes
SCA2	CAG	*ATXN2* (exon 1)	P	<31	31–32	32–200	Unknown
SCA3 (Machado–Joseph disease)	CAG	*ATXN3* (exon 8)	P	12–40	41–85	52–86	Unknown
SCA6	CAG	*CACNA1A* (exon 47)	P	<18	19	20–33	None found
SCA7	CAG	*ATXN7* (exon 3)	P	4–17	28–33	>36 to >460	Yes
SCA17	CAG	*TBP* (exon 3)	P > M	25–42	43–48	45–66	Yes
SMBA	CAG	*AR* (exon 1)	P	13–31	32–39	40	None found
Diseases with non-coding TNRs							
DM1	CTG	*DMPK* (3′ UTR)	M	5–37	37–50	<50	Yes
DM2	CCTG	*CNBP* (intron 1)	Uncertain	<30	31–74	75–11,000	Yes
FRAX-E	GCC	*AFF2* (5′ UTR)	M	4–39	40–200	>200	Unknown
FRDA	GAA	*FXN* (intron 1)	Recessive	5–30	31–100	70–1,000	Yes
FXS	CGG	*FMR1* (5′ UTR)	M	6–50	55–200	200–4,000	Yes
HDL2	CTG	*JPH3* (exon 2A)	M	6–27	29–35	36–57	Unknown
SCA8	CTG	*ATXN8OS* (3′ UTR)	M	15–34	34–89	89–250	Unknown
SCA10	ATTCT	*ATXN10* (intron 9)	M and P (smaller changes with M)	10–29	29–400	400–4,500	Yes
SCA12	CAG	*PPP2R2B* (5′ UTR)	M and P (more unstable with P)	7–28	28–66	66–78	None found

Are there other trinucleotide repeats of the sequence "CAG" that might be of interest to extend the study?

McMurray, C. *Mechanisms of trinucleotide repeat instability during human development*. Nature Reviews Genetics 2010 11: 786-99

The Table Browser

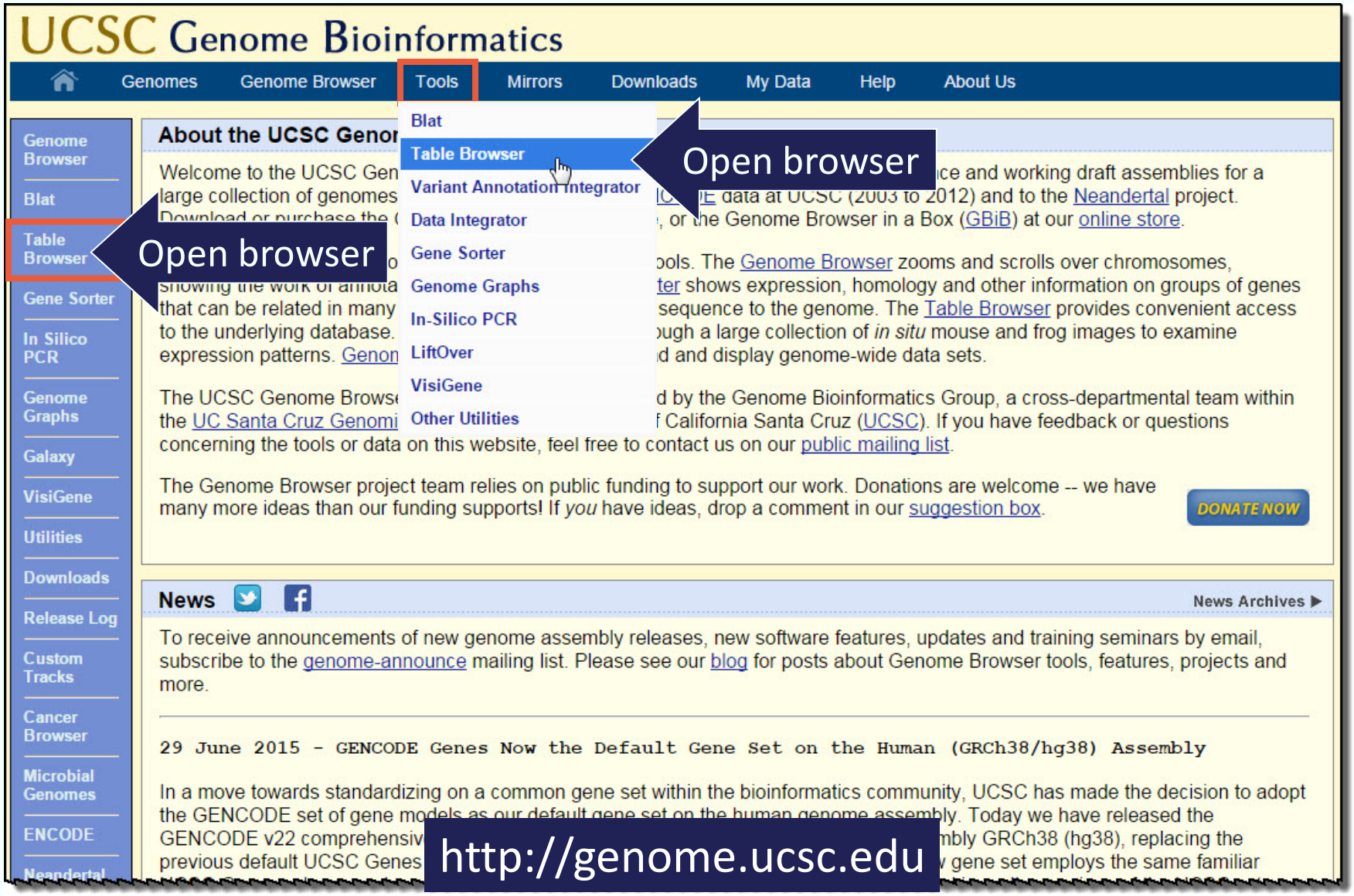
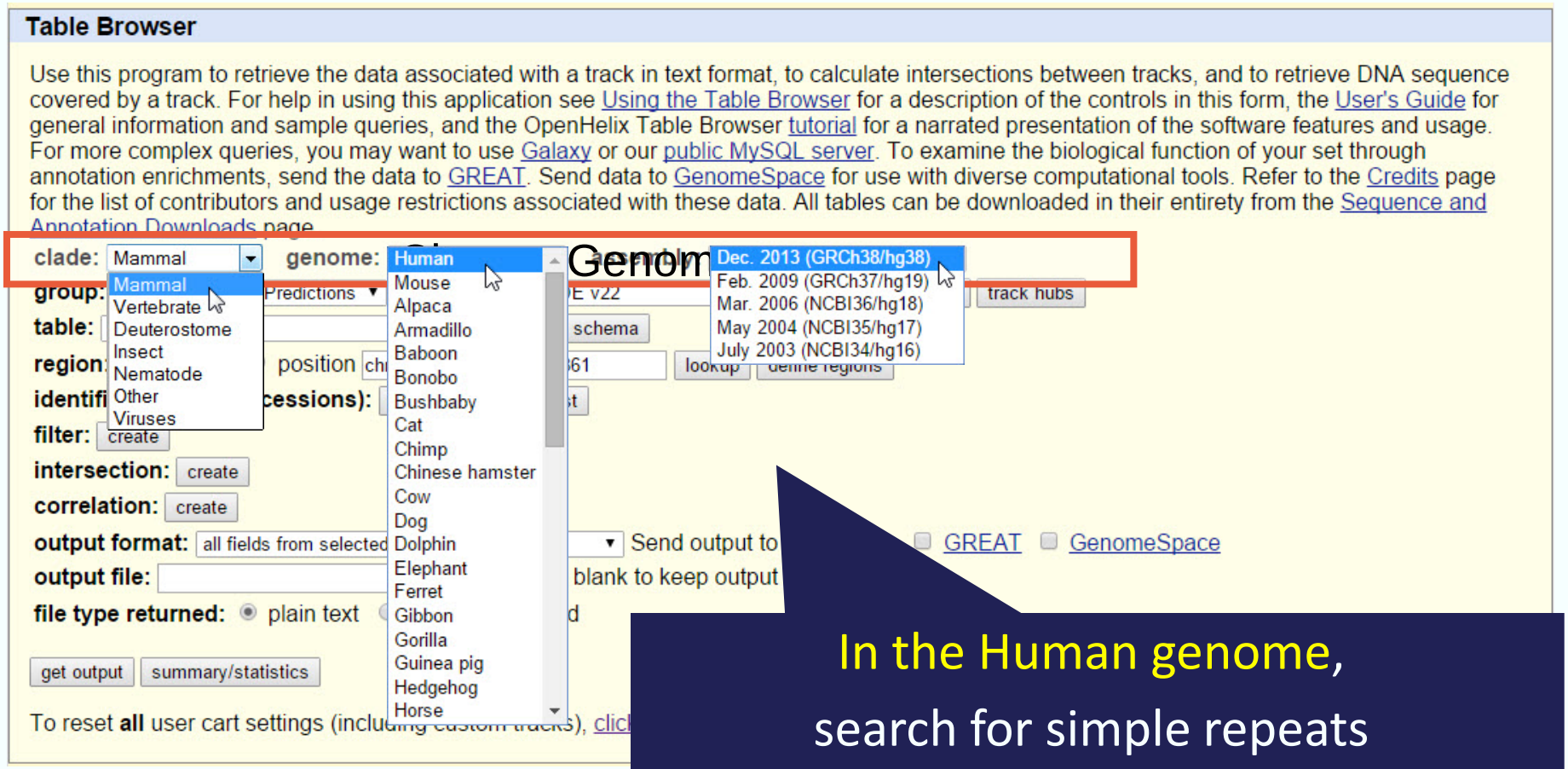


Table Browser: Choose Genome



In the Human genome,

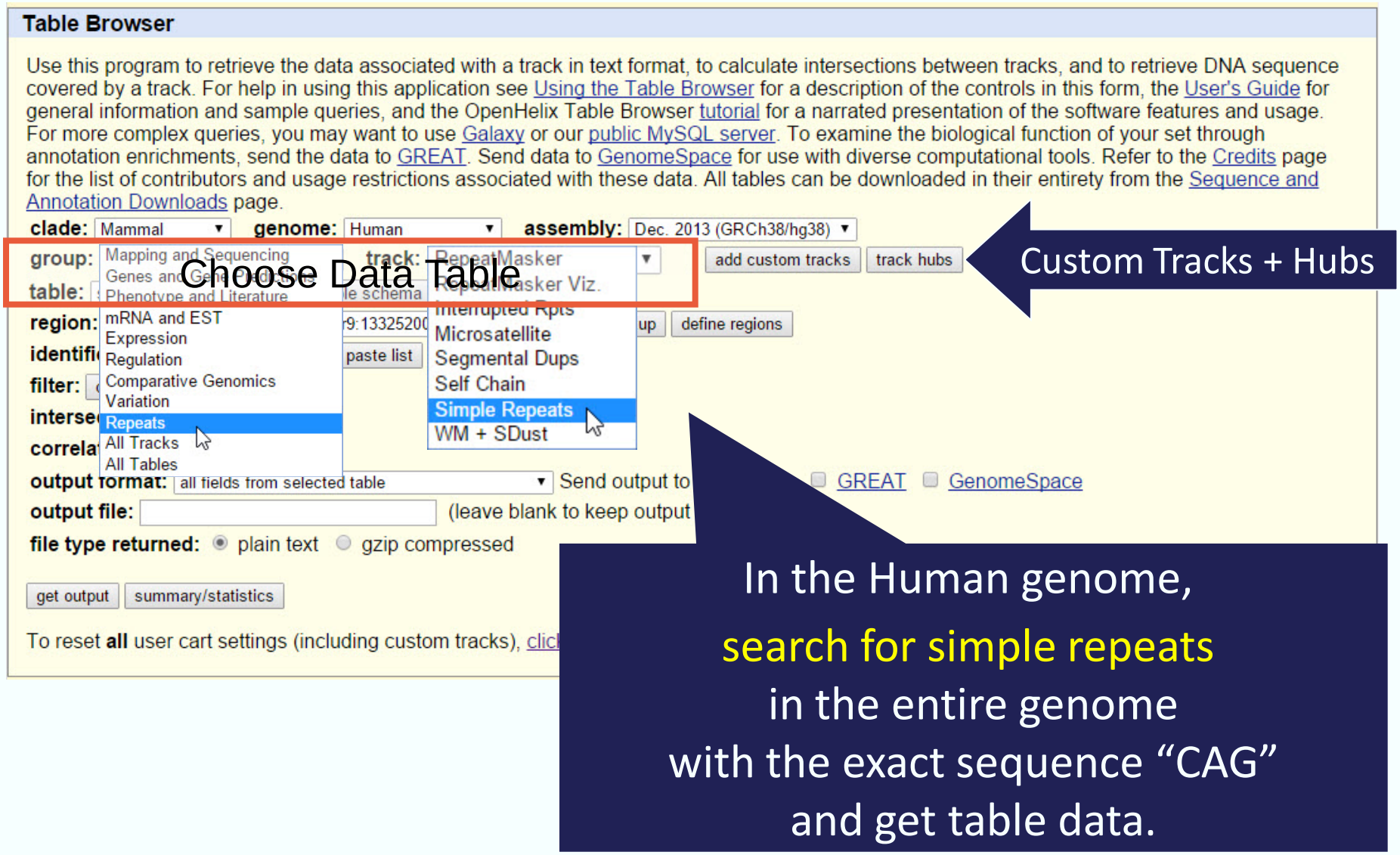
search for simple repeats

in the entire genome

with the exact sequence “CAG”

and get table data.

Table Browser: Choose Table to Search



Choose Data Table

Custom Tracks + Hubs

In the Human genome,
search for simple repeats
in the entire genome
with the exact sequence "CAG"
and get table data.

Table Browser: Describe Table

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Repeats **track:** Sim... Repeats add custom tracks track hubs

table: simpleRepeat describe table schema

Describe table

region: genome position chr9:133252000-13...0861 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Sen...

output file: (leave blank to ...

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click he...

Schema for Simple Repeats - Simple Tandem Repeats by TRF

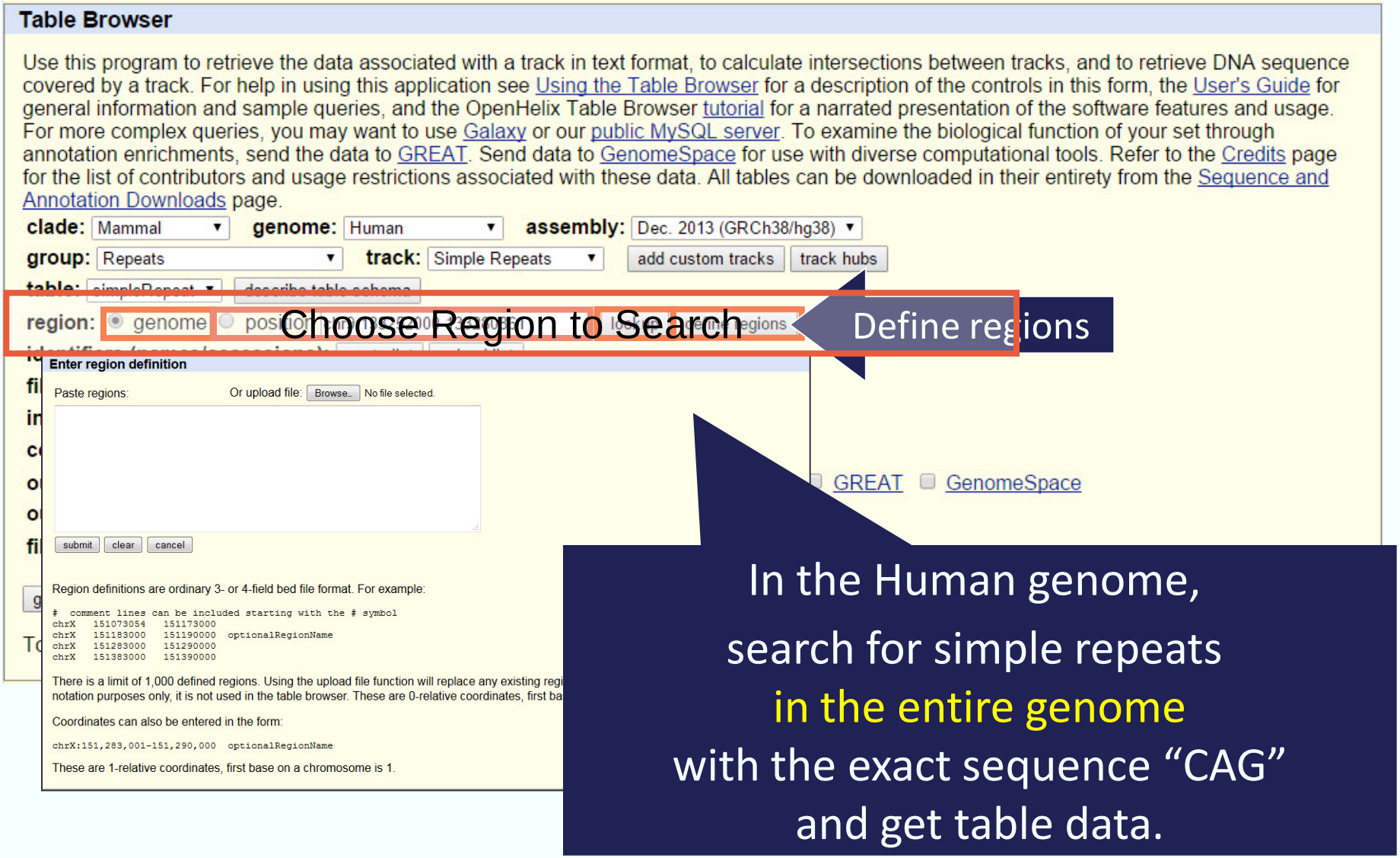
Database: hg38 **Primary Table:** simpleRepeat **Row Count:** 1,014,212 **Data last updated:** 2013-12-29
Format description: Describes the Simple Tandem Repeats

field	example	SQL type	description
bin	585	mallint(5) unsigne	Indexing field to speed chromosome range queries.
chrom	chr1	archar(255)	Reference sequence chromosome or scaffold
chromStart	10000	nt(10) unsigned	Start position in chromosome
chromEnd	10468	nt(10) unsigned	End position in chromosome
name	trf	archar(255)	Simple Repeats tag name
period	6	nt(10) unsigned	Length of repeat unit
copyNum	77.2	loat	Mean number of copies of repeat
consensusSiz	6	nt(10) unsigned	Length of consensus sequence
perMatch	95	nt(10) unsigned	Percentage Match
perIndel	3	nt(10) unsigned	Percentage Indel
score	789	nt(10) unsigned	Alignment Score = 2*match-7*mismatch-7*indel; minscore=50
A	33	nt(10) unsigned	Percent of A's in repeat unit
C	51	nt(10) unsigned	Percent of C's in repeat unit
G	0	nt(10) unsigned	Percent of G's in repeat unit
T	15	nt(10) unsigned	Percent of T's in repeat unit
entropy	1.43	loat	Entropy
sequence	TAACCC	ongblob	Sequence of repeat unit element

Sample Rows

bin	chrom	chromStart	chromEnd	name	period	copyNum	consensusSize	perMatch	perIndel	score	A	C	G	T	entropy	
585	chr1	10000	10468	trf	6	77.2	6	95	3	789	33	51	0	15	1.43	TAACCC
585	chr1	10627	10800	trf	29	6	29	100	0	346	13	38	47	0	1.43	AGGCGCGCCGCGC
585	chr1	10757	10997	trf	76	3.2	76	95	2	434	17	30	45	6	1.73	GGCGCAGGCGCAGA
585	chr1	11225	11447	trf	117	1.9	121	80	14	273	12	32	33	20	1.9	CGCCCCCTGCTGGC
585	chr1	11271	11448	trf	61	2.9	61	82	4	187	12	32	34	20	1.9	AGTGGTGGCACGCC
585	chr1	11283	11448	trf	62	2.7	61	82	2	199	12	33	33	20	1.9	CGCCCCCTGCTGGC

Table Browser: Choose Region to Search



In the Human genome,
search for simple repeats
in the entire genome
with the exact sequence “CAG”
and get table data.

Table Browser: Quick Summary of Expected Hits

OpenHelix
A Division of Bio-IT World

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2

group: Repeats **track:** Simple Repeats add cu

table: simpleRepeat describe table schema

region: genome position chr9:133252000-133280861 lookup

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Send output to

output file: (leave blank to keep outpu

file type returned: plain text gzip compressed

get output summary/statistics

Data Summary

To reset **all** user cart settings (incl ding custom tracks), click here.

Simple Repeats (simpleRepeat) Summary Statistics

item count	1,014,212
item bases	146,785,521 (4.8 1%)
item total	335,386,791 (11.0 %)
smallest item	25
average item	1
biggest item	500
smallest score	
average score	
biggest score	80

Region and Timing Statistic

region	genom
bases in region	3,209
bases in gaps	159
load time	
calculation time	
free memory time	
filter	
intersection	

~ 1 million Simple Repeats annotated in the Human genome

Table Browser: Filter to Refine Search

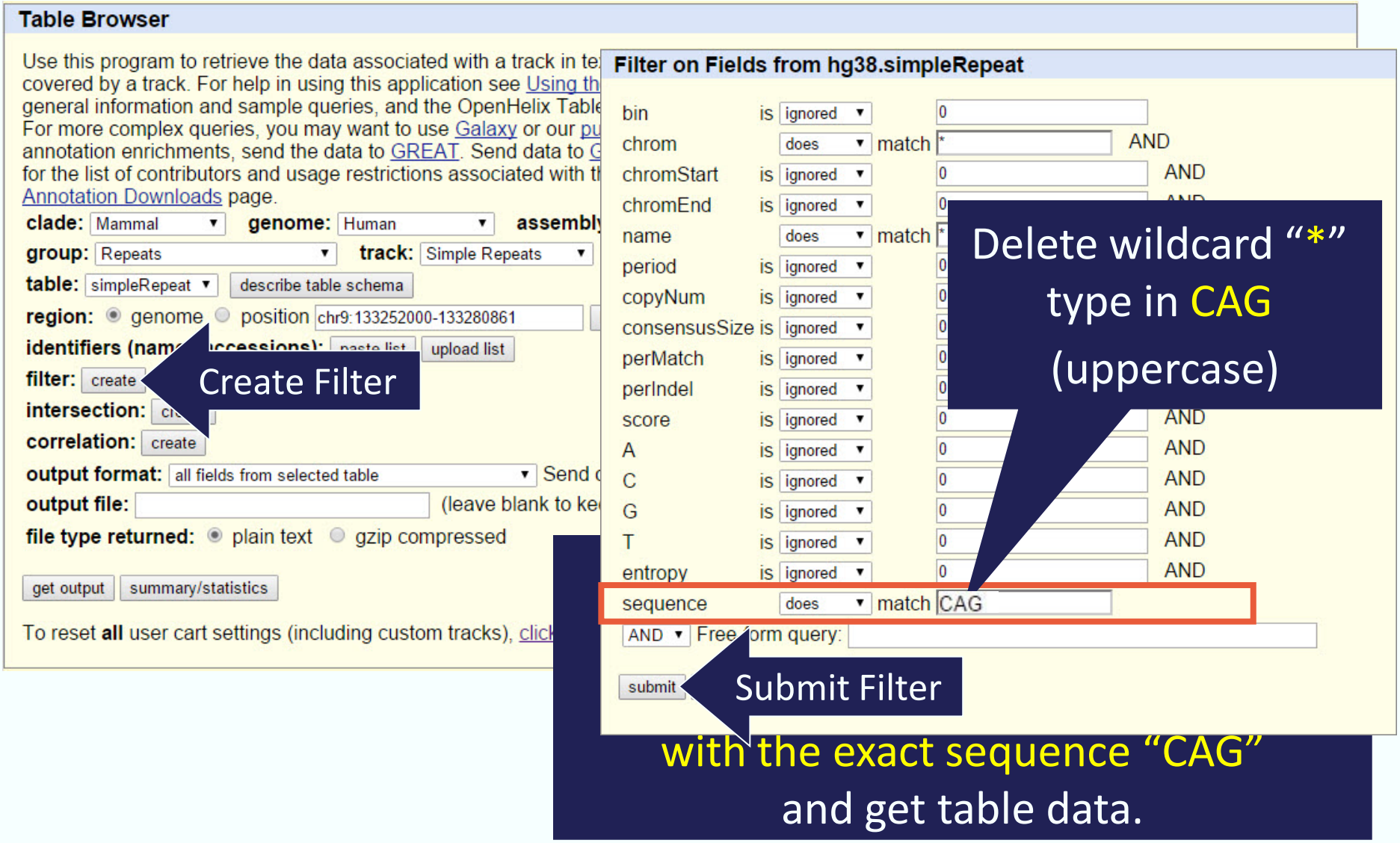
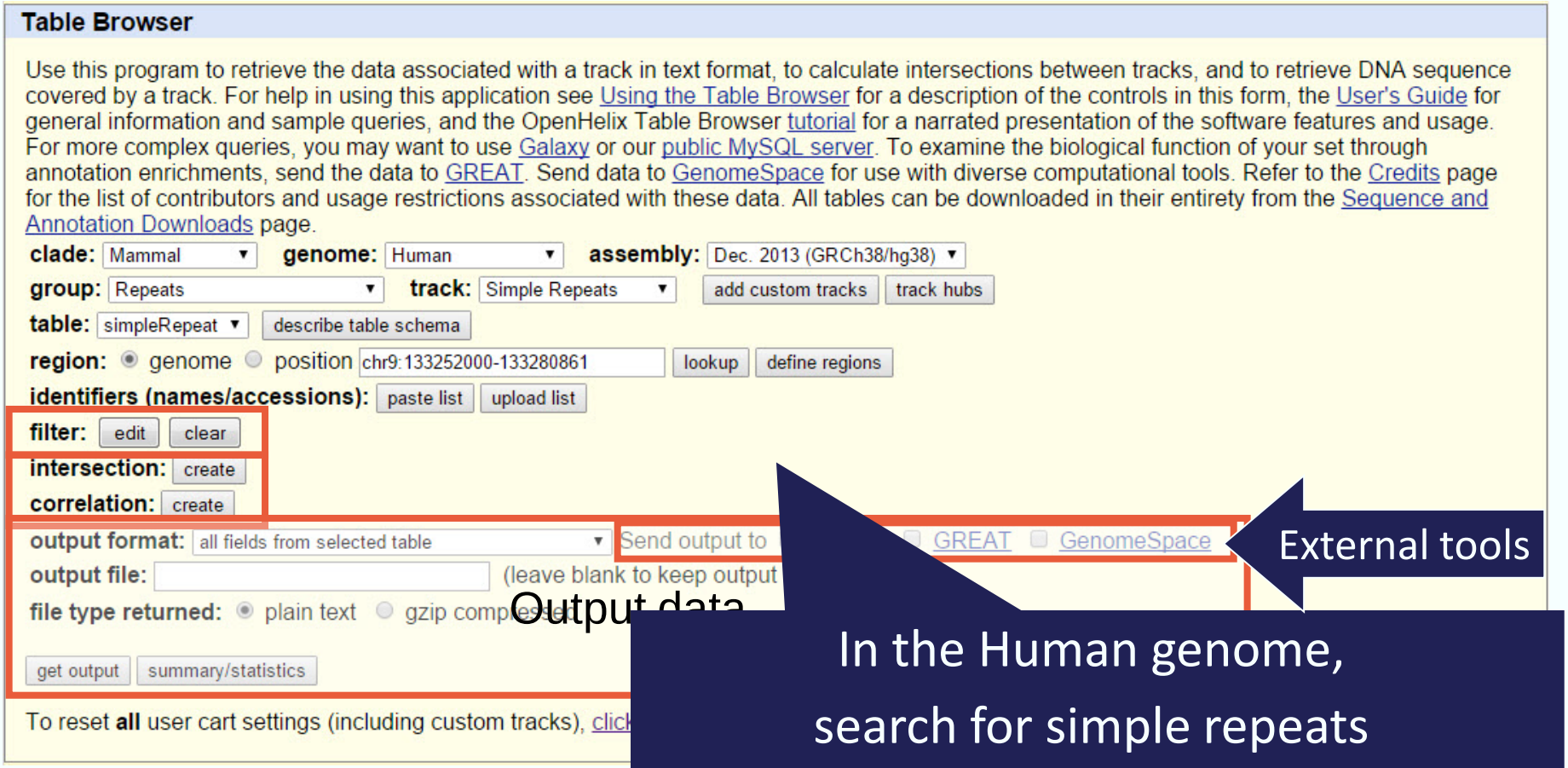


Table Browser: Output Data



In the Human genome,
search for simple repeats
in the entire genome
with the exact sequence "CAG"
and get table data.

Table Browser: Output Formats

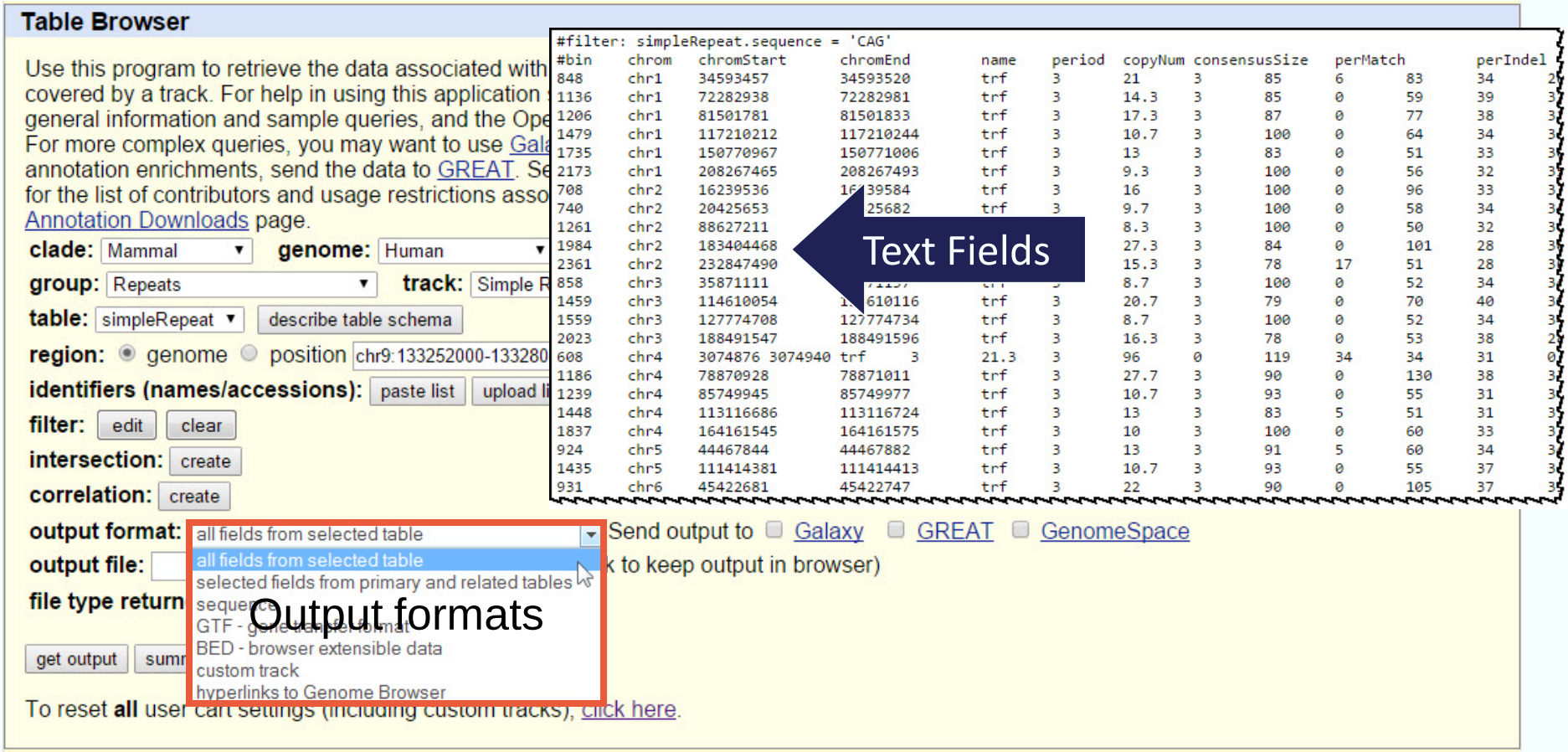


Table Browser: FASTA Sequence Output

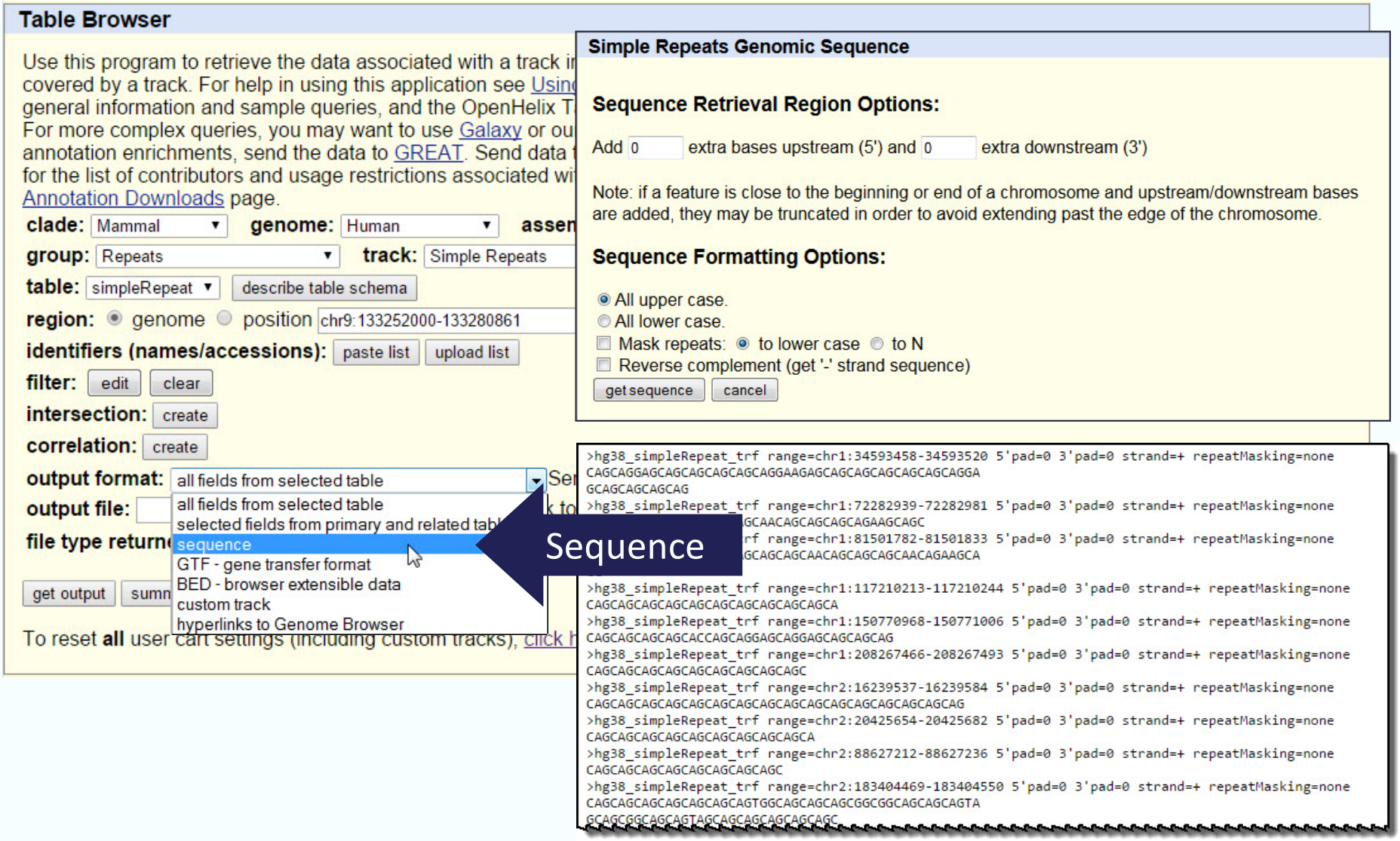


Table Browser: File Format Outputs

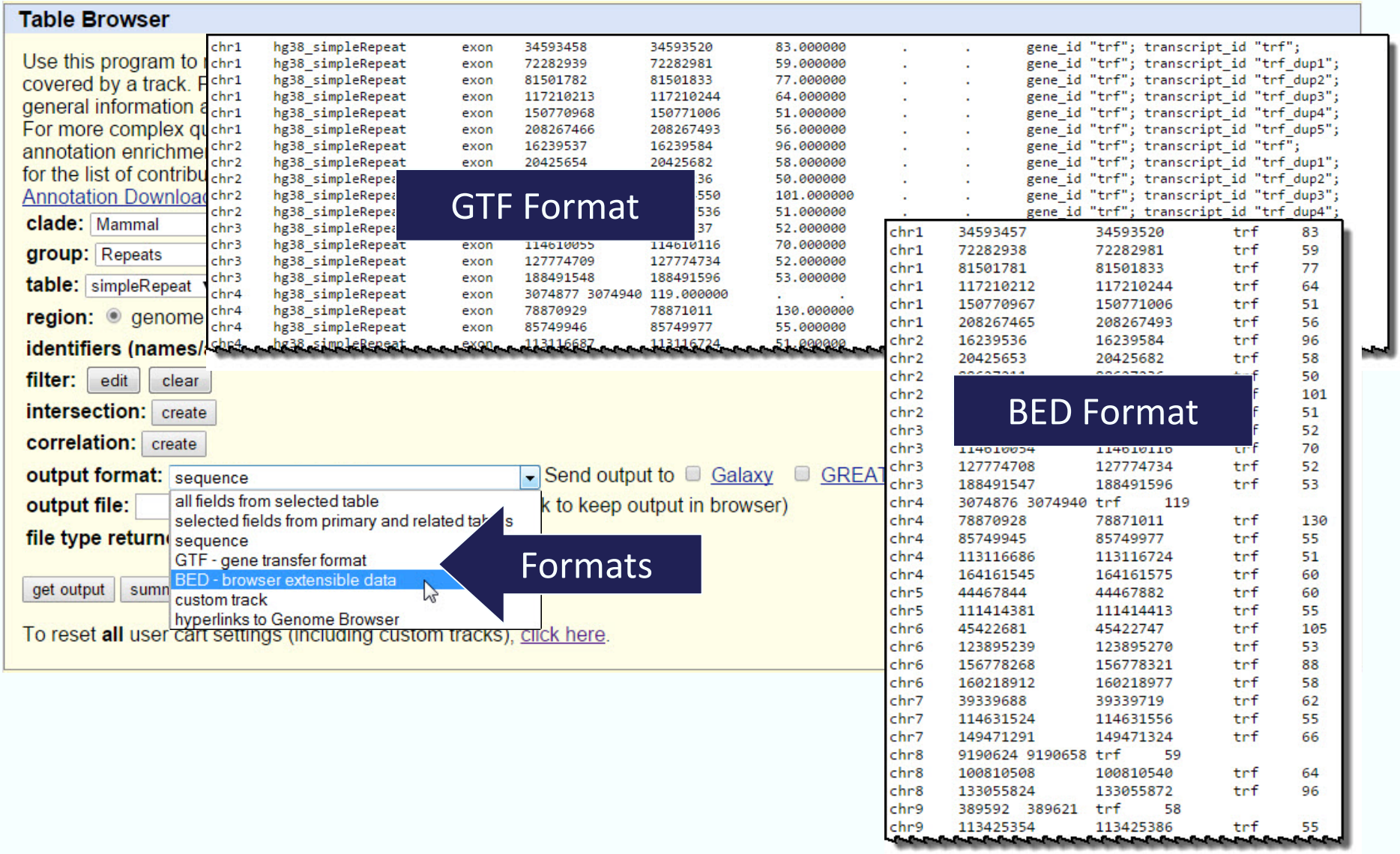


Table Browser: Custom Track Output

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human

group: Repeats **track:** Simple Re

table: simpleRepeat describe table schema

region: genome position chr9:133252000-1332808

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection: create

correlation: create

output format: all fields from selected table

output file:

file type returne

get output summ

- all fields from selected table
- selected fields from primary and related table
- sequence
- GTF - gene transfer format
- BED - browser extensible data
- custom track
- hyperlinks to Genome Browser

To reset **all** user cart settings (including custom trac

Custom Track

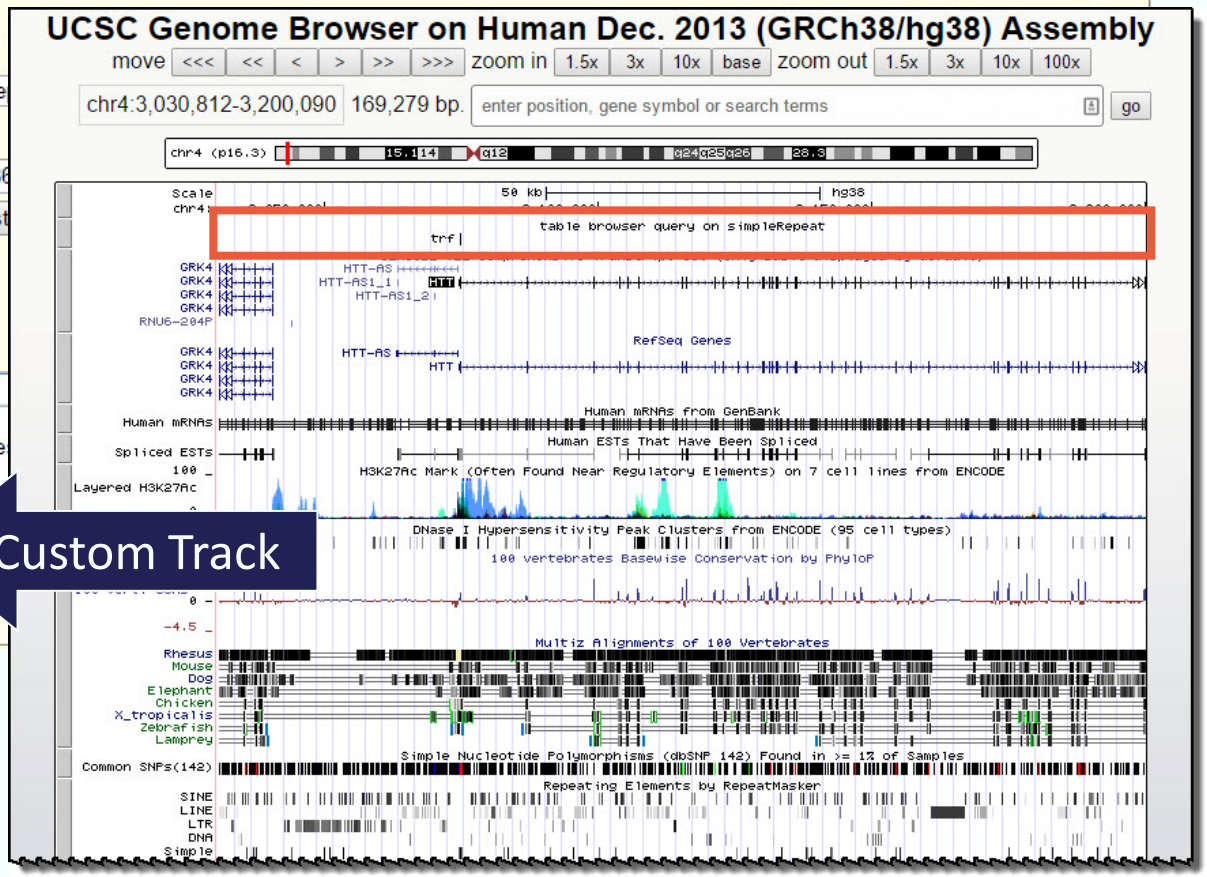


Table Browser: Hyperlinks Output

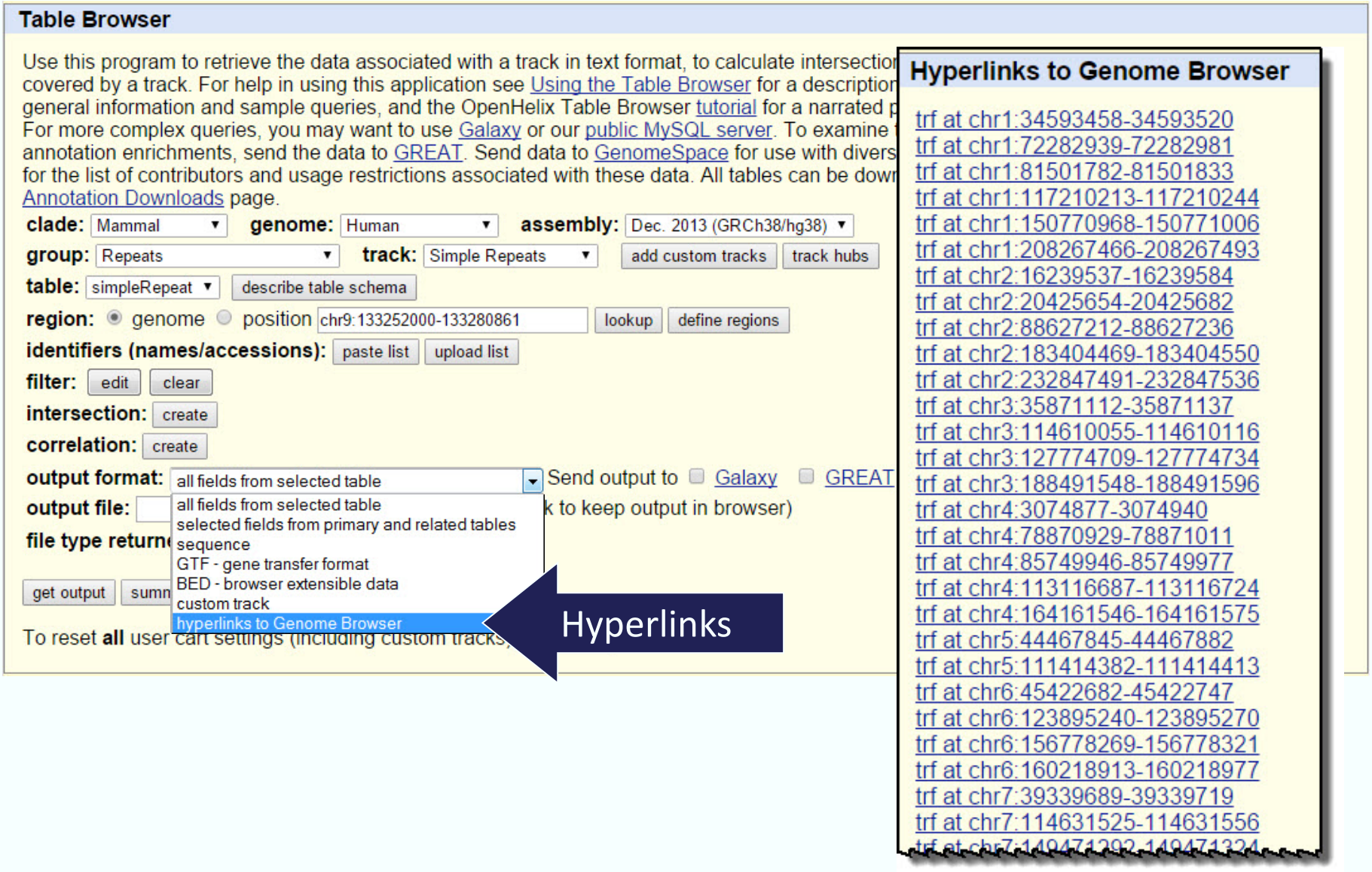


Table Browser: Obtaining Output

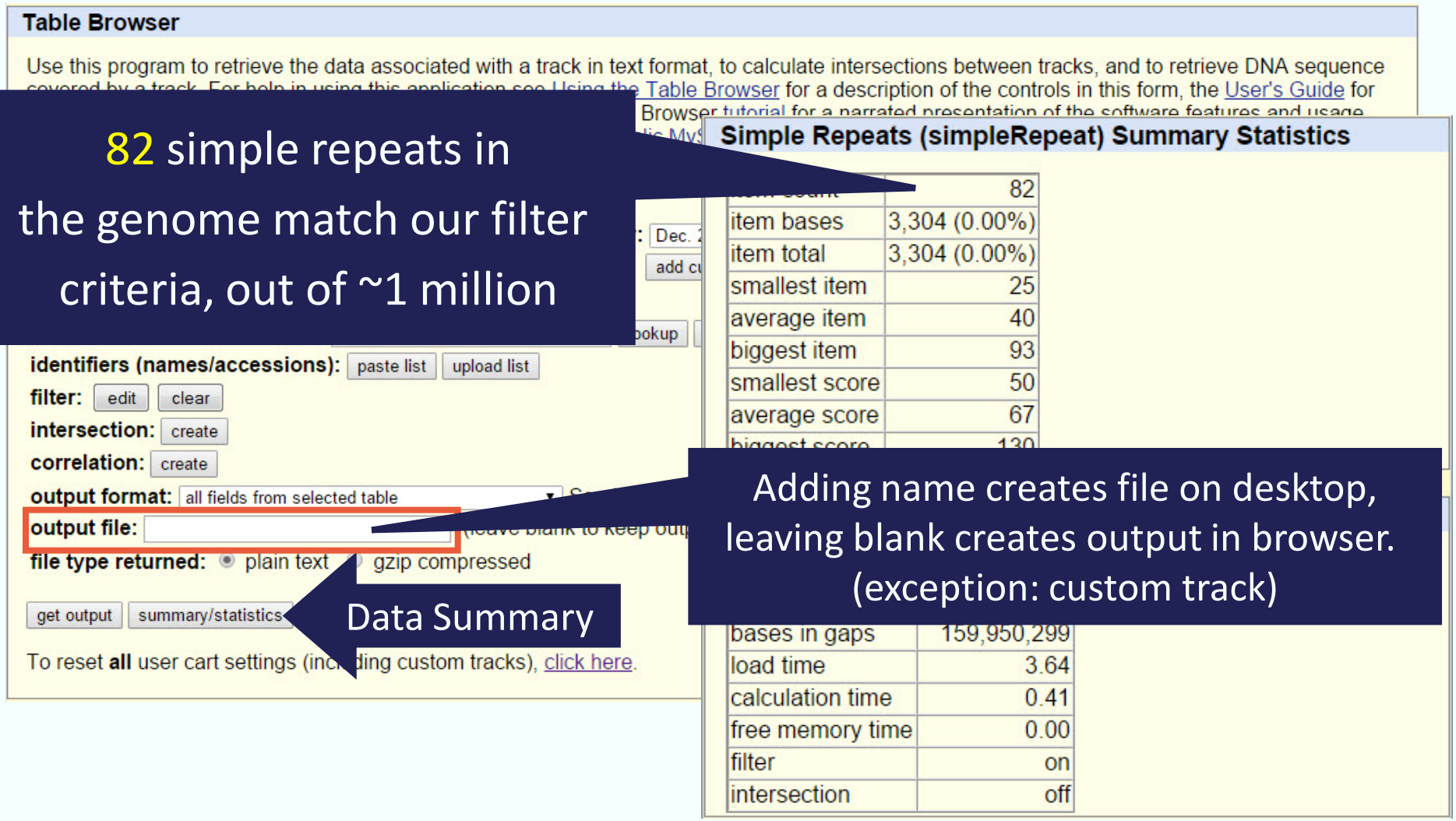


Table Browser: Output Configuration

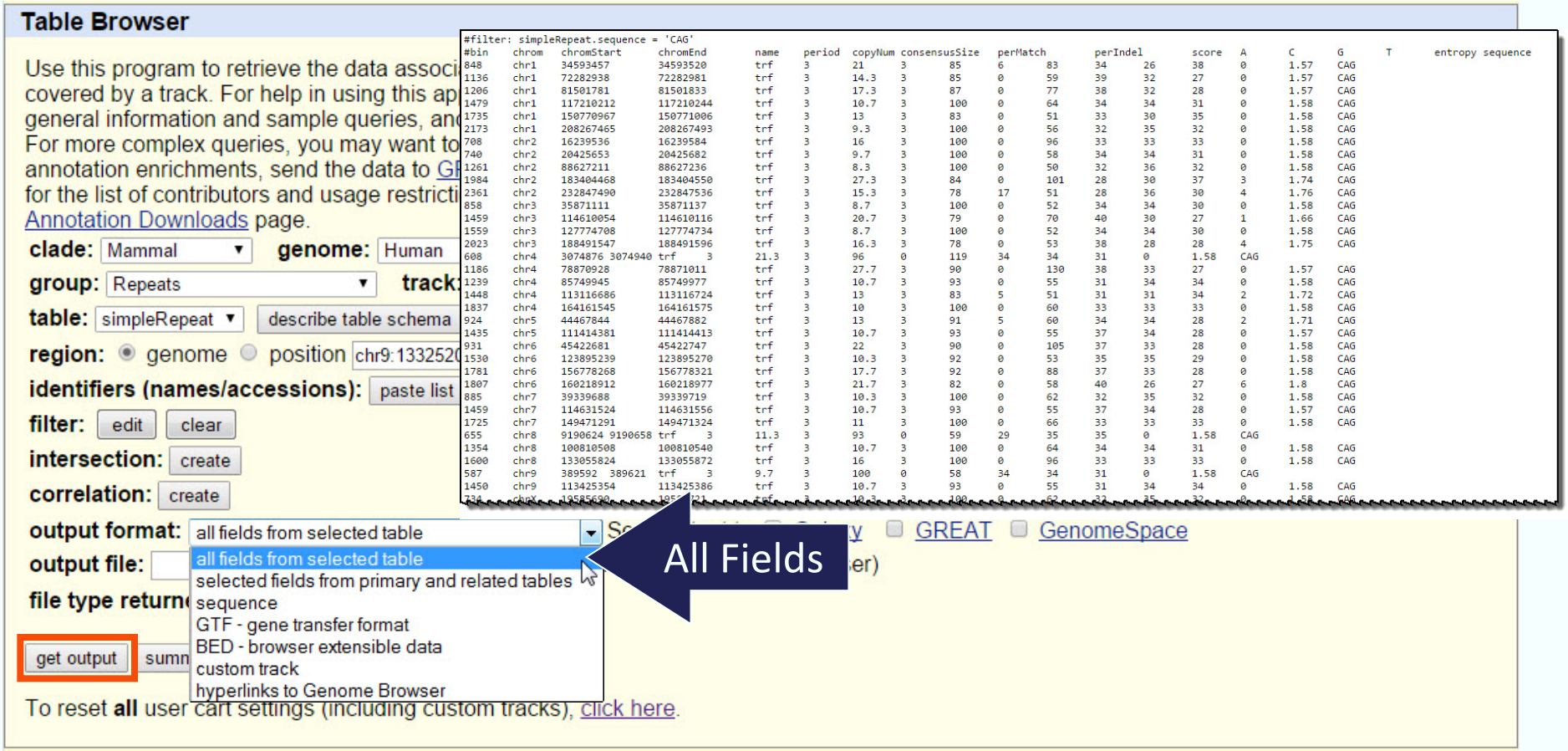
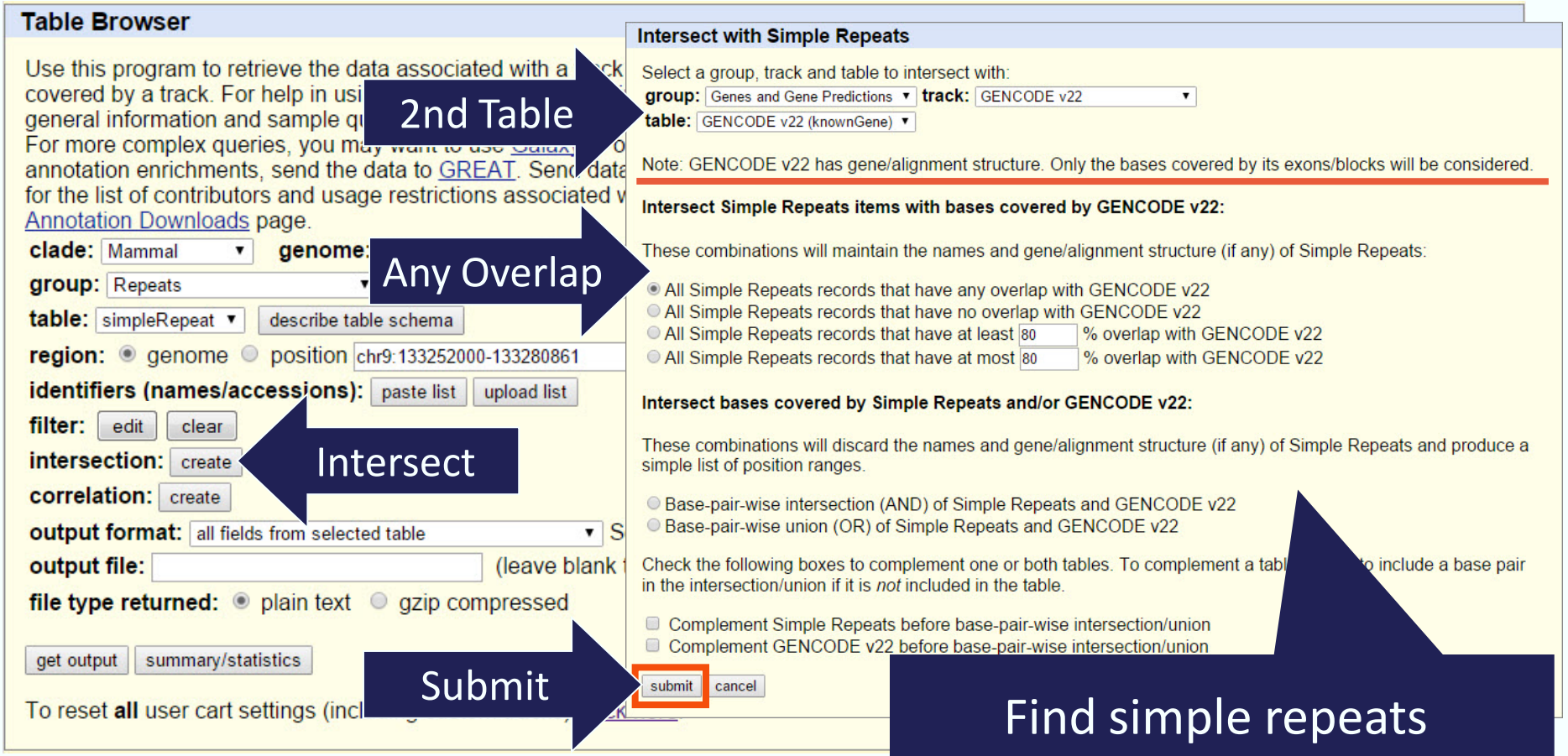


Table Browser and Custom Tracks Agenda

- Table Browser Introduction
- Table Browser: Set up and Filters
- Table Browser: Intersections
- Custom Tracks: Table Browser & User Generated
- Summary
- Exercises

UCSC Table Browser: http://genome.ucsc.edu

Table Browser: Intersecting Data



Find simple repeats (sequence = CAG) within known genes and get hyperlinks.

Table Browser: Intersecting Data Narrows Search

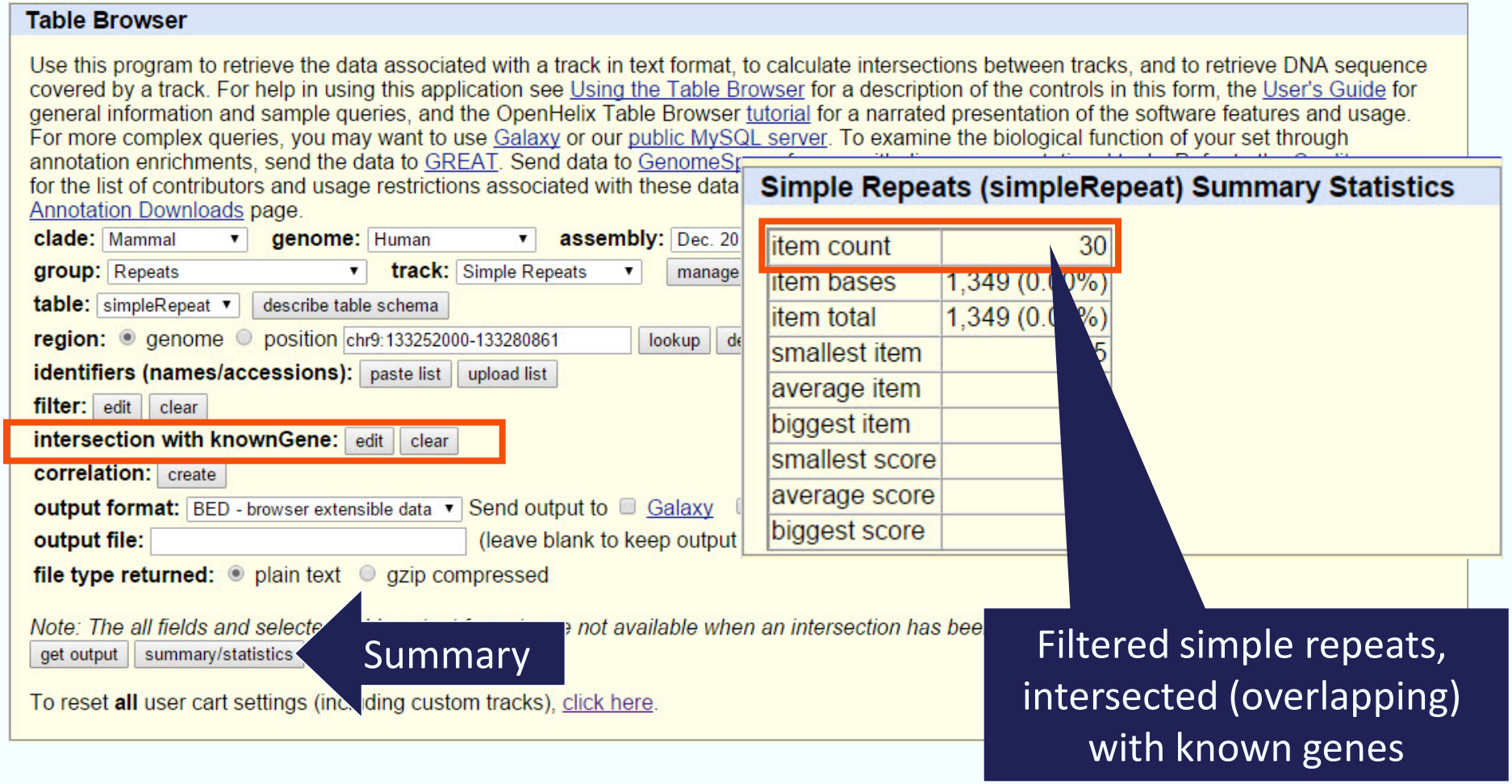


Table Browser: Get Hyperlinks

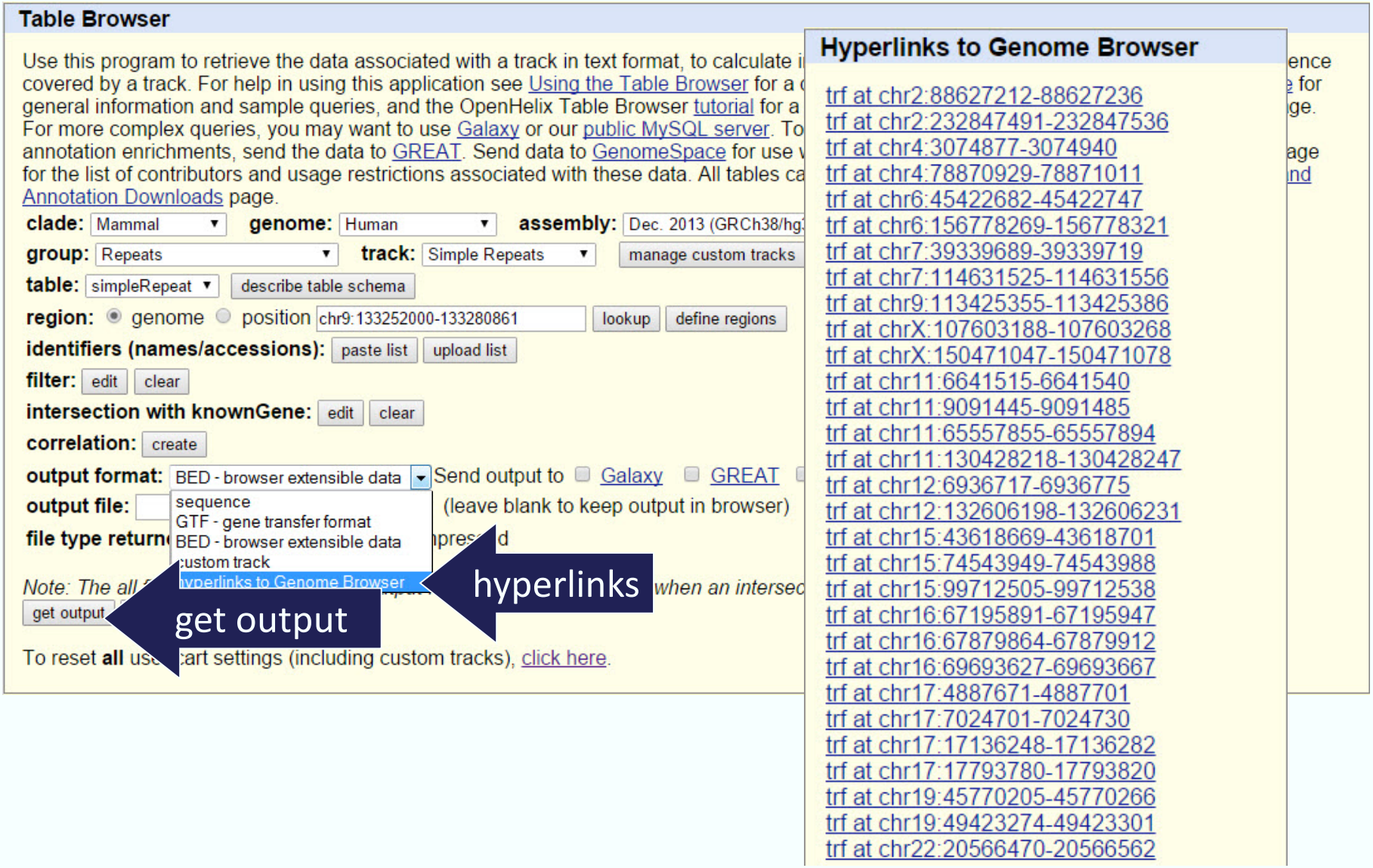


Table Browser: Exploring the results

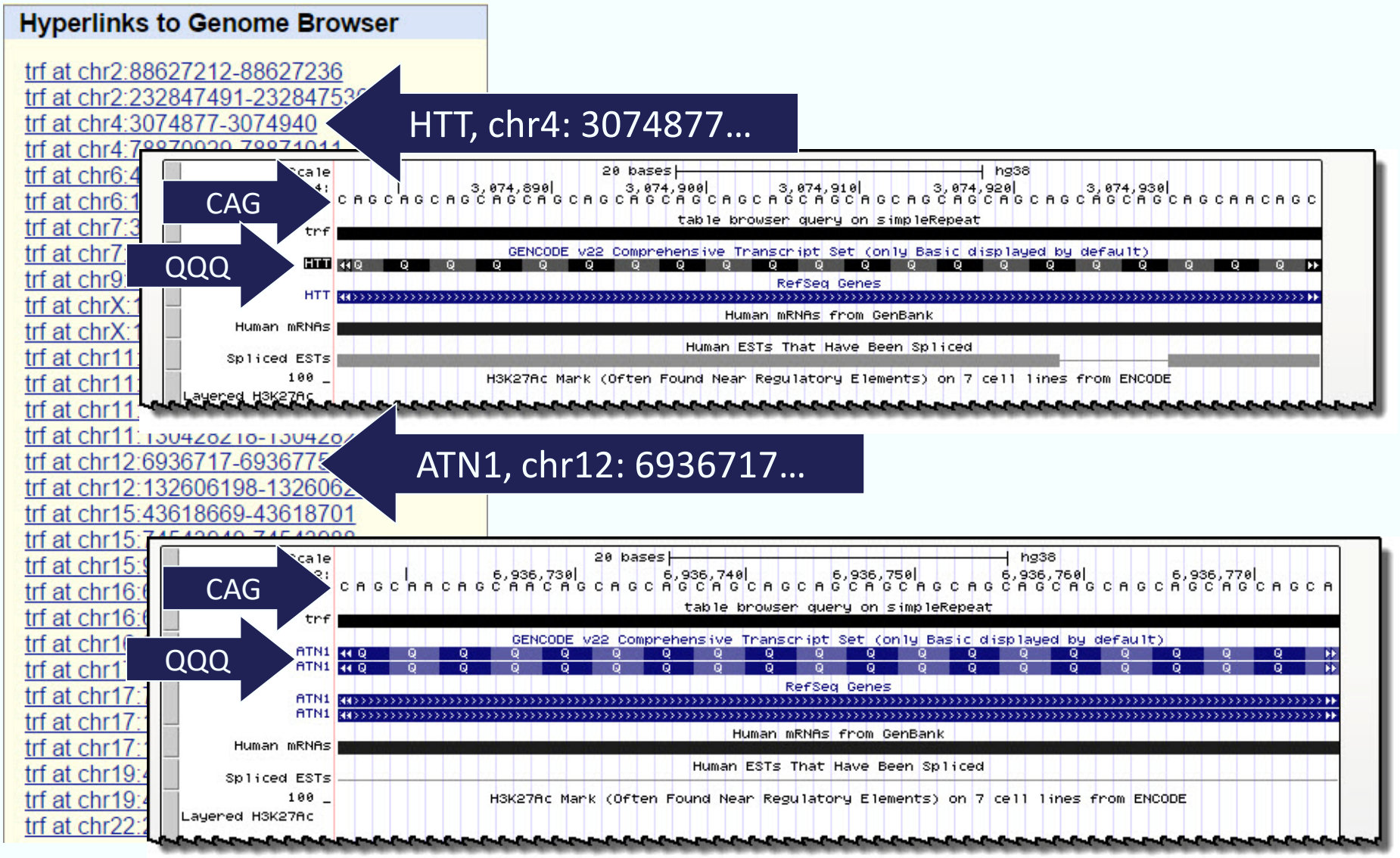
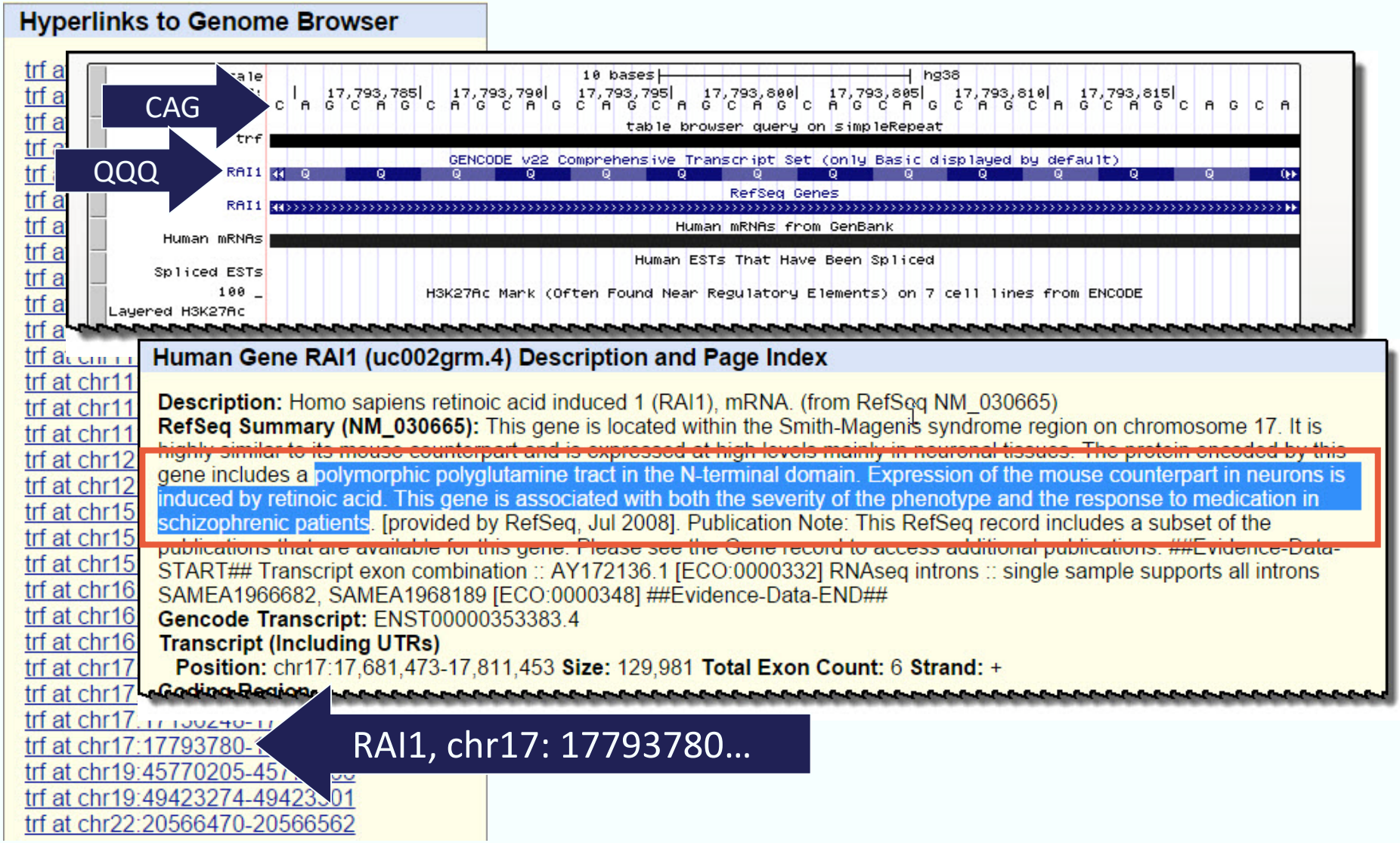


Table Browser: Exploring the results



Tutorial - UCSC Genome Browser

1) executar uma análise semelhante a descrita nos slides anteriores para identificar genes candidatos a doenças associadas a expansão de trinucleotídeos usando a ferramenta “Table Browser”

2) Anotação de novos RNAs não codificadores usando informações genômicas no UCSC Genome Browser

3) Integração com dados de bulk e single cell RNAseq e dados clinicos

Instruções para confecção do relatório:

Responder as perguntas de cada tutorial. Podem usar nas respostas as imagens geradas na analise.
O relatório deve ser entregue através do e-disciplinas.

Tutorial - Banco de dados

1. Anotação de novos RNAs não codificadores usando informações genômicas no UCSC Genome Browser

Em um estudo que reconstruiu o transcritoma de tumores pancreáticos (Paixão et al., 2022) foram identificados cinco RNAs não codificadores longos (lncRNAs) super-expressos que ainda não estão descritos no catálogo de referência de genes humanos (GENCODE)

A sequencia dos lncRNAs (em formato fasta) estão disponiveis na pagina da disciiplina (lncRNAs_upPDAC.fasta)

1) utilize o programa BLAT para alinhar as sequencias na versão mais atual do genoma humano (dica: escolha o melhor alinhamento, maior score, para prosseguir)
2) reporte as coordenadas de mapeamento dos transcritos incluindo: cromossomo, coordenadas de inicio e fim, orientação da fita do DNA.
3) reporte o numero de exons e o tamanho de cada transcrito maduro
4) reporte o nome e a distância (em bases) do gene mais próximo anotado na vizinhança dos RNAs.
5) Ative tracks com informações de elementos regulatórios da expressão gênica (ex. promotores, enhancers, ilhas CpG) e busque evidências que corroborem a transcrição do lncRNA candidato. Reporte se encontrou elementos regulatórios na proximidade do inicio de transcrição de algum dos RNAs.
6) Ative tracks com informações de conservação evolutiva e busque evidências que corroborem a existência de seleção da sequencia dos lncRNAs candidatos. Reporte se observou trechos conservados nas sequencia de algum dos RNAs.

2. Integração com dados de bulk e single cell RNAseq e dados clinicos

Na mesma análise foram identificados vinte lncRNAs já anotados no GENCODE, alguns dos quais com expressao alterada já descrita em tumores de pancreas ou outros contextos tumorais. Para avançar no estudo desses RNAs é importante saber em que tipo celular são expressos (vários tipos celulares no microambiente tumoral). Bancos de dados de expressão gênica que integram dados de “bulk tissue” (RNAseq) e células únicas (scRNAseq) possibilitam acesso a dados experimentais já anotados.

Os identificadores dos RNAs (gene symbol) estão na tabela abaixo.

LINC01559	AC005550.3	RP11-350J20.12	MIR210HG
LINC02577	CCAT1	UNC5B-AS1	LINC01614
RP11-460N11.2	RP11-284F21.10	UCA1	RP11-38M8.1
LINC00675	RP3-340N1.2	MMP12	ACER3
LINC01133	RP11-284F21.9	RP11-395B7.4	LINC00941

1. No portal Cancer Single-cell Expression Map https://ngdc.cncb.ac.cn/cancerscem/ busque por informações sobre a expressão desses lncRNAs em estudos de single cell e bulk tissue RNA seq.
2. Para quantos desses RNAs foi possivel encontrar informações?
3. Algum desses RNAs tem um padrão de expressão que sugira ser preferencialmemte expresso em tumores de pâncreas? Nota: os projetos com amostras de pancreas tem o nome de PDAC ou TCGA-PAAD
4. No portal Gepia2 http://gepia2.cancer-pku.cn/ verifique se a expressão de algum dos RNAs acima está correlacionada com a sobrevida global (overal survival) ou sobrevida livre da doença.